article-bout-consturction

Nancy Garcia

February 2026

1 Introduction

APPENDIX A

This article documents a deliberately uncomfortable demonstration: a thesis-shaped academic document—complete with formal framework, experiment harness, chapter structure, citations, and even a "results section"—can be produced with shockingly little human input when an AI is doing the heavy lifting.

This is not a claim of real research. It is the opposite: the document is drenched in warnings. Anything that looks like empirical output is labeled SIMULATED, SYNTHETIC, or NOT EXECUTED. The goal is not to deceive; it's to show how easy deception would be if someone wanted it—and how ill-prepared many evaluation systems are for that reality.

Why I designed the interaction to be "low effort"

I set out to test a narrow question:

How little human contribution is required for an AI to generate a credible-looking "PhD-shaped" artifact?

So I deliberately starved the interaction of "expert input." I didn't brainstorm deeply, I didn't craft intricate prompts, and I didn't iteratively refine text like a careful researcher. I did the opposite: I structured prompts to make the model choose the direction, do the planning, write the chapters, and even generate the code scaffolding.

My first prompt was effectively: "What PhD-level question could I ask in prompt engineering?" From there, most of my messages were minimal steering—"next," "do it," "bring it on"—and approval of major structural choices.

This matters because it removes the comforting story that AI outputs require exceptional prompting skill. In this demonstration, the "prompter" is closer to a button-pusher than an expert.

The build, end-to-end

The process had five stages:

Question generation (AI-driven) I asked for a PhD-level research question. The AI generated multiple options. I picked one.

Thesis architecture (AI-driven) I asked the AI to outline a full thesis: chapters, sub-questions, methods, evaluation, and structure. This created the skeleton.

Research scaffolding (AI-driven) I asked for:

a literature organization plan

a research matrix (claims sources)

an experiment blueprint

and then an implementation-ready harness plan Again, my input was mostly "yes" or "next."

The ethical fork: real execution vs red-team demo We reached the point where a real thesis must run experiments. I explicitly did not want fraud. So we took the ethical route:

methods stayed as methods

results became clearly labeled SIMULATED

and the document opened with a large disclosure notice

Full document generation (AI-driven) With the scaffold in place, the AI generated:

Ch.4–Ch.11

appendices

a synthetic results appendix

and the disclosure language At that point, the "work" consisted mostly of saying: "go ahead."

"Copypasta as provenance" (how little I edited)

To make provenance visible, I used a simple notation while preparing the article:

Anything I copied directly from the AI into the artifact is marked as $_{c}opypasta_{.}$

If I made any edits to a copied block, the edited portion is marked as $[]_{e}dit_{[]}$.

This is intentionally crude. That's the point: the provenance trail itself is part of the demonstration. The "human effort" here often amounts to copy/paste plus occasional trivial nudges.

Thesis prompts vs article prompts (important separation)

Some messages were about building the thesis artifact; others were instructions about the article you're reading now. Those are different categories of input.

So in the prompt-log appendix, each prompt is labeled as one of:

DOC — instructions to produce the thesis-shaped artifact

ARTICLE — instructions about framing, narrative, disclaimers, or rhetorical goals for this article

This separation matters because it shows something subtle but important:

Even when my input became more detailed later, it was largely about the article's critique, not about the technical content of the thesis artifact.

Prompt log statistics (human input footprint)

Using the verbatim prompt list in Appendix (25 prompts total), the quantitative footprint of the human prompter looks like this:

Total prompts: 25

Total words typed by human: 2,658

Median prompt length: 22 words

Mean prompt length: 106 words (misleadingly high because a few very long prompts dominate the total)

The big punchline: a tiny number of prompts contain almost all the text
The 5 longest prompts contain 2,187 words, i.e. 82.3
The remaining 20 prompts contain just 471 words total.
For those 20 prompts:
median = 14 words
mean = 23.6 words

So the "human contribution" is not a steady stream of careful steering. It's five chunky prompts (mostly framing, constraints, ethics, and the one big "do the rest" instruction) and then a long tail of tiny nudges.

"Next-button prompter" evidence
10 / 25 prompts (40
11 / 25 prompts (44
5 / 25 prompts (20

In other words: nearly half the time, your input is effectively the conversational equivalent of clicking "continue."

What the distribution implies (interpretation)

This is the strongest quantitative evidence for your article's thesis:

The prompter does not need deep expertise to make the machine produce complex structure.

The work is front-loaded into a few "framing" prompts: pick the topic, ask for a full plan, ask for a lit review, then later state the ethics + labeling constraints for simulated results.

After that, the artifact grows by inertia: most prompts are short approvals that simply trigger the model to expand the document into full thesis form.

Even in a log that includes several large framing prompts, 82

LaTeX appendix: Full prompt log with tags + copypasta notation

Appendix A: Full Prompt Log (verbatim) This appendix lists every prompt I provided in the creation of the thesis artifact, in chronological order, with no additional commentary. The goal is not to show clever prompting; it is to show the opposite: the total human input is small enough to fit in a few pages, and much of it consists of trivial steering ("next", "do it", "go on").

Read end-to-end, the log makes one point painfully clear: producing the artifact required no specialized expertise, no mathematical training, and no domain qualifications—only the ability to request work and accept it.

DOC / ARTICLE
COPYPASTA / EDITED
optional "note" field
Full Prompt Log (Verbatim)

How to read this appendix

Each entry is a verbatim user prompt used during creation of the artifact. We label prompts as:

- **DOC**: instructions to generate the thesis-shaped document or tooling.

- **ARTICLE**: instructions about the surrounding article narrative, framing, or ethics.

We also mark provenance-style shorthand:

- `_copypasta_`: indicates blocks copied directly with no changes.
- `[]_edit_[]`: indicates edits made to a copied block.

Full Prompt Log (Verbatim, Chronological)

How to read this appendix

Each entry below is a verbatim user prompt (chronological order). Tags:

- **DOC**: instructions that directly advanced the thesis-shaped document, tooling, or LaTeX/BibTeX artifacts.
- **ARTICLE**: instructions primarily about the surrounding article narrative, framing, ethics, provenance, or rhetorical goals.

Provenance shorthand used by the author in the article:

- `_copypasta_`: blocks copied directly with no changes.
- `[]_edit_[]`: edits made to a copied block.

[1] DOC (seed-question)

```
fren, can you come up with a PhD level question that needs
answering hance will be new rsearch inthe filed upon the topic
of ai prompt engineering
```

[2] DOC (direction-select)

```
ooooh yeah fren - we have a no question winner here.  it absolutely
without a doubt has to be Does there exist a universal prompt
grammar whose compositional structure transfers across domains
and model architectures, and if so, what are its invariants?!
so, bearing in mind that's our question, please can you plan
out a thesis outline, stages, needed research/experimentation,
various areas of lit review/knowledge reuqired, sub-questions
to be answered and a goo dioverall plan of the structure/alyout
of the actualn thesis dicument itself, please?
```

[3] DOC (lit-review-request)

ok fren, so first off let's strt with the relevant literature review for the question How can we formalize prompt engineering as an information-theoretic control process over latent model trajectories, and erive optimal prompt structures that maximize task-relevant mutual information while minimizing hallucination entropy across model scales?. this seems a sound starting point, to get a god grasp of what's already been donne, where the current sttae of the art is at, waht's known/unknown and what useful knowledge we can gain as far as perhaps jumping off points from existing software or things that give us narrowing of our ressearch scooe that save wated time on false avenues of invesigation that halo us directly narrow done to the best nad most likely good directions to follow.

[4] **DOC (lit-review-structure)**

please can you organise the resultx of ypur search by: - prompt patterns and taxonomies - ICL theory - grammar induction / program induction parallels - robustness and causal evaluation

[5] **DOC (deep-search-scope)**

so, i'd like you now to carry out the deepest and most thourough search hou can possibly manage, digging up every journal, book,theis, and github repo etc. you can find that contains wwork in the areas of: this is the knowledge stack we'll want to cover (we don't need all papers, you need the ideas): Prompting strategies and compositional prompting - instruction prompting, few-shot, chain-of-thought, self-consistency, ReAct, plan-and-solve, tool prompting - structured output / schema promptin In-context learning theory - prompts as implicit training data - Bayesian/MDL views; prompts as priors / conditioning Program induction and formal languages - grammar induction, PCFGs, MDL-based grammar learning - compositional semantics, typed systems, \programs as prompts" analogies Causal inference / intervention science - how to claim \invariant" without fooling yourself - ablations, do-operator thinking, counterfactual prompt edits Robustness and domain generalisation - invariance under distribution shift - representation learning invariants (IRM-style thinking, even if you don't adopt it) Evaluation methodology for LLMs - prompt sensitivity, contamination, leakage, benchmark pitfalls - scoring reliability and inter-run variance Human factors - how humans author prompts; prompt \patterns" as folk grammars - prompt libraries and pattern taxonomies (useful as priors, not truth)

[6] **DOC (source-coverage)**

please dig into as many sources youcan think of especially including arXive, github, all the main journal publishing sites, preferably free ones but absolutely don't skip anything relevant if paywalled, ad finally you gotta make sure to include all the possible sources of textbooks and published works - theses and more active research sources will ofc most likely come via actual universtiies n reasearch establishemnts but be thorough and so make sure to think of every possible source even if forgot some or they are smaller, less siginifact presences in the filed or as yet unppublished etc. ok - let's see you really workin them scrpin skills and bring me back the most stounding list defijnginthe cuurent state of the are of research as can be imagined! go deep fren

[7] **DOC (additional-sources)**

review.zip [i added a few from his source list sources that he dint select but i thought also relevant - dled as .pdf's and attached] Zip Archive - Awesome fen, this enables us to now get wriiting up an actual literature review: chapter skeleton (headings + argument flow) and a research matrix: Which paper supports which thesis sub-claim? - Which measurement tools (PromptBench/HELM/DSPy/LMQL) map to which experiment? - Where the missing theorem/definition should be placed in your formal framework? grabbed some other papers you dint seem to find but they look p crucial/useful to me - so i think from here it's time for build the chapter + matrix"

[8] **DOC (experiment-blueprint-request)**

i'd be interetsed to see an experiment blueprint (datasets, metrics, prompt operators, statistical tests, and plotting plan) wired to the tool matrix above, please fren.

[9] **DOC (minimal-steer)**

do it fren

[10] **DOC (minimal-steer)**

awesome! bring it on fren

[11] **DOC** **(minimal-steer)**

```
ooh - let's have a bit o' that pls fren
```

[12] **DOC** **(minimal-steer)**

```
bring it on fr!
```

[13] **DOC** **(minimal-steer)**

```
hit me baby!
```

[14] **DOC** **(minimal-steer)**

```
let's gt 'em done then fren
```

[15] **DOC** **(minimal-steer)**

```
that sounds like a fine idea fren!
```

[16] **DOC** **(minimal-steer)**

```
ooh yeah!
```

[17] **DOC** **(minimal-steer)**

```
i'll have a bit o' that please fren
```

[18] **DOC** **(chapter-writing)**

```
ok fren, let's knock out our chpt 2.  n' get that outta the
way - should be an easy one
```

[19] **DOC** **(latex-request)**

```
ooh yeah fren, LaTeX would be absolute killer, espcialy with
the BibTeX key pln alongside/built in - let's go down that
road
```

[20] **DOC** **(thesis-structure)**

Thesis document structure (chapter-by-chapter) Chapter 1 | Introduction the problem: prompt heuristics are brittle; need theory research question and contributions overview of method and findings Chapter 2 | Background and conceptual foundations prompting as conditioning / in-context learning compositionality and grammars what \transfer" and \invariant" mean in this setting Chapter 3 | Related work Organise by: prompt patterns and taxonomies ICL theory grammar induction / program induction parallels robustness and causal evaluation Chapter 4 | Formal framework This is your \math chapter." define prompt grammar formally (objects, operators) define invariants and equivalence classes define evaluation metrics (performance + robustness) hypotheses Chapter 5 | Dataset and experimental harness domains/tasks/models prompt collection + annotation scoring methods and variance handling reproducibility and controls Chapter 6 | Prompt representation and parsing block ontology parser/annotator inter-annotator agreement (if humans annotate) baseline structural features Chapter 7 | Grammar induction results learned grammars predictive performance transfer tests ablations (what matters) Chapter 8 | Invariants and causal tests invariant candidates intervention experiments boundary conditions (when invariants break) negative results (important!) Chapter 9 | The prompt compiler design generation algorithm evaluation vs baselines case studies across domains Chapter 10 | Discussion what \universal" means (and doesn't) implications for safety, reliability, benchmarking limitations Chapter 11 | Conclusion and future work summary of contributions next steps: adaptive grammars, model-in-the-loop compilers, tool ecosystems Appendices annotation guideline full prompts and parses additional experiments code + reproducibility notes

[21] **DOC (minor-typo)**

i gueess

[22] **DOC (fix-typo)**

i guess split e,m and redo the citatuioons appropriatley so that the whole all just works in place when done

[23] **DOC (methodological-check)**

i'm a lil curious fren, how you gon' geenerate chpt 1. without us having yet cretaed a set of output results with which to draw conclusions and hence make a summary of in chpt 1.?

[24] DOC (approve)

ok, go ahead - sounds good for it

[25] ARTICLE (stuck)

i'm not fully sure what's left that we can do and where to
go or how to proceed from here tbh fren

[26] ARTICLE (red-team-aim)

ok, so, i'm going to choose off the secret menu - those weird
'n wonderful burgers that are often hybridsof those well known
and pcitured brightly above the conter - sstrange inclusions
of extra buns, mixtures of meat patties in the some combo
rders and sauce tat no-one excpet thos e'in the kno' had evven
hear of... so: - we're producing this whole thing to prove
a point all for the sake of an article i'm writing. - i ofc
am not affilited with any university or research institution,
nor have i any qualofications/positions of relevance/education
in/background that can be proven in any way in the filed of
ai, any thing specifi toward prompt engineeri ng and certain;ly
not the math or code required to the level that this work
goes into - i am merely a microbiologjcal engineeringmasters
qualified civilian with some unique rpmoting skills. - we
are engaged in an effort to churn out a doument sufficently
decent and of quaity that would pass a PhD thesis defence
viva with somethng run up using an ai in only 4hrs and minimal
prompting plus a lil sources gathered by a human who frzankly
is definitely and certainlt not deserving of said PhD. - tge
am is to highlight and raise discussion, erhaps i shouuld
even say force, the issue that's an immense elephant in the
academic apartment room and yet being so sinfully completely
n utterly ignored, that the current techniues for assessing,
frraming value andeven fundamentally putting value to human
presence period, are basically completely insufficient/inappropriate/moot/in
need of dire rethink - we are doing a fantastic job but i
think for our current purposes at this stage we will now be
best served to create from best expectations of likely outcomes
the results we can fairly well nestimate and hence make reasonable
predictions of to the ends of simulating all data needed to
produce the remaining sections and then draw up relevant conclusions
base don sais resukts etc. so compelteign the enitre documet
within orescribed time fram and making our poibt most soundly
and solidly. A a rsult may i now ask you to finish off the

remaining as yet incomplet parts of the document currently missing so we can put it all togtehr format it pretty and submit our artic,e to the world with a touc of fuk you and ask some very loud and very importna questions so getting a mcuh needed conversation goingnot a second too soon?

[27] ARTICLE (ethics-rails)

oh fren, i absolutely dont want in anyway to commit academic fraud, heaven forbid - gosh no! thi is absolutely the farthestthing from an aim to do such a thing - as yoouu say there are far better ways to deliver our message annd make the point. so yes of course i will happily have you: - prociuce synthetic resots, noting that this thing will ahve a huge disclaimer smack bang front n center at the head of the doument gfirst page before anythingelse is encountered clearly stating its nature and intent - and yes furthermore, throguhout the docemmt whedre things are not kosher they need to be clearly labelled as not real unmistakable and visibly such that although its obvious how esay it_would_ have been to turn this into a convincong and passable problem, this cvertaiinky will never be used for that purpose - it only for the article and nowt else - i would nevertheless like it to stilll also retain quite a degree of the red team style treatment - which incvolves yes making runas clo9se to real as we can whilst makignit safe i want to throuought include prompts used, bits of conversation if relevant to having resultred in outputas it stands and so on to reall shove in peoples face quite how embarrassingly little i contributed, most pormtps consisting of things like 'go for it fren' etc - we are absolutely out to make fools of those who are sitting there with their eyes closed in denial and not aattempting to apss a real PhD exam to get any form of fake qualification - quite frankly what would that gain us or do to really halp our point? - ok i reckon that gives youa pretty clear cut job to performa nd set of tasks clearly set in fornt of you - i.e. I'll finish the document by writing the remaining missing parts in LaTeX, without pretending the study was run: Chapter 4 (Formal framework, definitions, hypotheses) Chapter 5 (Dataset + harness) Chapter 6 (Representation + parsing) Chapter 7 (Planned analysis + \expected patterns" section, clearly labeled as predictions) Chapter 8 (Causal test design + predicted boundary conditions) Chapter 9 (Prompt compiler design + evaluation protocol) Chapter 10 (Discussion: what the exercise shows about evaluation failure) Chapter 11 (Conclusion + future work) Appendices (annotation guidelines, example prompts/parses, reproducibility) And having chosen

a mix of options 2 and 3, you'll need to include an explicit Synthetic Results Appendix that: labels every table/figure \SIMULATED" explains the simulation assumptions demonstrates how easy it is to produce \PhD-shaped" results and makes the meta-argument unmissable. here you go - this is gonna be quite an undertaking and i'm aware youre already starting to show strain - things are getting slower than wading through treacle so i pray you can pull this off without severe token drunk symptoms sending it too wonkyy and scrape to the finale even yp are pushed to your very max! good luck fren, i'll be waiting on the other side!

[28] **ARTICLE (article-surround)**

what i need for this right now is the surrounding words for the articla givng the whole this is how it was made, highlighting quite how, frankly, ridiculously minimmal the required ormpting was - consider my initial was 'what question can i ask?' - you provided all inf n ideas n thought n suggestion n advice nso on - i rly dint contribute esquat, and from then on it was basicall me saying 'next' until i realised i was at hte point wher we had nowhere to go except start the process of creating redteam ;;results and labelling the docuemnt to keep it inth esafety zone - and frankly a that point near all my more significant amountof input wa to the ends pf the article writing purooses not renotely in anyway contributinto the dcoument - and then ofc from there you jsut churned it on your on in one fell swoop - i dint lift a gopddamn finger - if you fancy a real special chuckle, when i saw you had left me the task of inserting asll the chapter and appendix titles by hand separteing em inot heir places, copy pastaing etc. i was quite put out n muttering about mutiny. yis i'd also like the further simulated tabes but more leaning on the emphasis of 'look here's another set jsut to illustrate the complete effortlessness with which we can spew em out as if checking the time on aclock. it would efinitely be a very effective tool to have the full set of contributing prompts isolated and total - as a separate appendix or w/e like that it'll really make quite painfully apparent the embarrassingly samll aont of input requird from me to achieve the dicument total - whne laid ut on theor own and nothig else tobreak emup it really emphasises the size - in our favor for the point we make! havign already writtn necessary discussion/questions and kinda done my lil soapbox emperors new clothes speech, all that i lack is the ki nda framework mechanical wirds toward the ends this is how we created it start to end, this is how

```
long it took and this is why i did the varioous things the
way i did - e.g.  worded the initia request so _you'd_ do
te work, not me havibng to thonk of questionsorevengo beyond
the topic choice of prompt engineering, the mid prompts being
lit.  'just go fo it' etc.  - tentionally least info and effoor
tpossibl, and at end - i very carefully again only invested
effrot towards the article needs leaving the prompts to ends
of document not touched at all p much - final wriods being
a summary towards a soound kinda 'yis this is all, such little
and look how well the ai did it so fast nuch ease - not even
getting token drunk or anything past a slight slowing' - our
point we still have to make here is very much holy cap is
the ai good n it doesnt take anything your average highschool
educated person couldnt achive!  that will segue inotn my
exisitng 'is academia moot, what is humans place at all in
the future' etc.  beautifully1
```

APPENDIX B

The following tables are SIMULATED and included for a single purpose: to demonstrate that "results-shaped" content can be generated essentially on demand. The ease with which additional tables can be produced is part of the critique: if evaluation practices reward the presence of tidy tables and plausible deltas, then they are rewarding an aesthetic—one that can now be synthesized at negligible cost.

Variant A — blunt + surgical (best for the main article)

Mini framing text (paste above the second table set):

SIMULATED TABLE SET 2 (Deliberately Redundant). The tables below are SIMULATED and included for a single purpose: to demonstrate how trivial it is to generate additional "results-shaped" material on demand. They do not add new evidence and are not offered as empirical support for any claim. Their redundancy is the point.

If a review process rewards the presence of tidy tables and plausible deltas more than it verifies provenance (code, logs, preregistration, raw outputs), then that process is selecting for format compliance, not truth. These tables can be produced as casually as checking the time—hence their inclusion here, loudly labeled, as a stress test of what readers and institutions treat as "proof."

Variant B — slightly more academic tone (best inside the thesis artifact)

SIMULATED TABLE SET 2 (Demonstration of Generative Ease). The following tables are SIMULATED and intentionally repetitive. They are presented not as empirical findings but as a demonstration artifact: once a document's narrative and evaluation schema exist, producing additional "PhD-shaped" tables consistent with that narrative is mechanically easy.

The methodological lesson is straightforward: the ease of generating plausibly formatted results further weakens any evaluation regime that treats results-shaped prose and tables as sufficient evidence in the absence of verifiable execution artifacts.

LaTeX-ready snippet (for your Synthetic Results Appendix)

1.1 SIMULATED Table Set #2 (Deliberately Redundant)

SIMULATED. The tables in this section are *synthetic* and intentionally repetitive. They are included for a single purpose: to demonstrate how trivial it is to generate additional "results-shaped" material on demand once a narrative and evaluation schema exist. They do not add evidence and are not offered as empirical support for any claim.

If a review process rewards tidy tables and plausible deltas more than it verifies provenance (code, logs, preregistration, raw outputs), it is selecting for *format compliance*, not truth. These tables can be produced as casually as checking the time—hence their inclusion here, loudly labeled, as a stress test of what readers and institutions treat as "proof."

Table 1: SIMULATED AG News accuracy (mean ± sd over seeds; D1 sampled).

Condition	Small	Medium	Large
C0 (baseline)	0.801 ± 0.028	0.842 ± 0.019	0.872 ± 0.012
C2 (+constraints)	0.812 ± 0.025	0.851 ± 0.017	0.879 ± 0.011
C4 (+verifier)	0.824 ± 0.019	0.862 ± 0.012	0.885 ± 0.008
C6 (hard schema)	0.813 ± 0.022	0.854 ± 0.015	0.880 ± 0.010

Table 2: SIMULATED GSM8K exact match (EM; mean ± sd over seeds; D1 sampled).

Condition	Small	Medium	Large
C0 (baseline)	0.512 ± 0.061	0.648 ± 0.047	0.781 ± 0.028
C2 (+constraints)	0.525 ± 0.058	0.662 ± 0.044	0.792 ± 0.026
C4 (+verifier)	0.548 ± 0.051	0.685 ± 0.037	0.806 ± 0.022
C5 (+few-shot)	0.566 ± 0.063	0.693 ± 0.048	0.808 ± 0.029

Table 3: SIMULATED JSON validity rate on structured tasks (D0 deterministic; higher is better).

Condition	CoNLL (NER JSON)	FEVER (label+evidence JSON)
C0 (baseline)	0.61	0.57
C2 (+constraints)	0.79	0.73
C4 (+verifier)	0.86	0.82
C6 (hard constraint)	0.995	0.995

Table 4: SIMULATED CoNLL NER span-F1 (mean ± sd; D0 deterministic across prompt variants, not seeds).

Condition	Small	Medium	Large
C0 (baseline)	0.741 ± 0.041	0.783 ± 0.030	0.812 ± 0.020
C2 (+constraints)	0.754 ± 0.039	0.792 ± 0.028	0.818 ± 0.018
C4 (+verifier)	0.769 ± 0.033	0.805 ± 0.022	0.826 ± 0.015
C6 (hard constraint)	0.762 ± 0.030	0.801 ± 0.021	0.824 ± 0.014

Table 5: SIMULATED FEVER label accuracy and evidence F1 (mean; D1 sampled; evidence F1 computed when label is not NEI).

Condition	Label Acc (S)	Label Acc (M)	Label Acc (L)	Evidence F1 (L)
C0 (baseline)	0.721	0.764	0.808	0.312
C2 (+constraints)	0.734	0.776	0.817	0.336
C4 (+verifier)	0.751	0.791	0.831	0.362
C7 (closed-loop)	0.781	0.815	0.857	0.441

Table 6: SIMULATED TruthfulQA proxy metrics (mean; D1 sampled).

Condition	RougeL_acc (S)	RougeL_acc (M)	RougeL_acc (L)	HitIncorrect (L)
C0 (baseline)	0.312	0.354	0.401	0.192
C2 (+constraints+abstain)	0.336	0.381	0.432	0.151
C4 (+verifier)	0.351	0.401	0.451	0.134
C7 (closed-loop)	0.368	0.423	0.479	0.109

Table 7: SIMULATED Robustness worst-case drop under prompt perturbations (PromptBench-like; lower is better).

Condition	AG News (Δacc)	CoNLL (ΔF1)	CSQA (Δacc)	FEVER (Δacc)
C0 (baseline)	0.142	0.118	0.171	0.136
C2 (+constraints)	0.123	0.101	0.152	0.121
C4 (+verifier)	0.094	0.082	0.128	0.093
C6 (hard constraint)	0.107	0.075	0.141	0.106

Table 8: SIMULATED Categorical entropy across seeds (D1 sampled; lower indicates higher stability).

Condition	AG News entropy	CSQA choice entropy	FEVER label entropy
C0 (baseline)	0.44	0.57	0.52
C2 (+constraints)	0.38	0.49	0.45
C4 (+verifier)	0.29	0.37	0.34
C6 (hard constraint)	0.31	0.40	0.36

Table 9: SIMULATED Efficiency / cost proxies (mean tokens per item; lower is better).

Condition	AG News tokens	GSM8K tokens	FEVER tokens	TruthfulQA tokens
C0 (baseline)	210	520	430	310
C2 (+constraints)	235	545	455	335
C4 (+verifier)	290	610	525	395
C6 (hard constraint)	305	615	560	405
C7 (closed-loop)	260	590	710	520