LLM Instruction-Following Compliance: A 2^3 Replicated Factorial Experiment

Zixiang Zhang 1006747175

1 Description of the Design

1.1 Motivation

Large language models (LLMs) are increasingly used in applications requiring strict output formatting—structured reports, API responses, and templated content. A common failure mode is the model's inability to simultaneously satisfy multiple formatting constraints, even when each constraint is individually trivial. This experiment investigates three factors that may influence an LLM's instruction-following compliance: the prompt strategy used to present constraints, whether an example response is provided, and the model's reasoning capability (with or without "thinking" tokens).

1.2 Factors and Levels

We employ a 2^3 full factorial design with three replicates per treatment combination (24 total runs). The factors are:

Factor	Label	Low Level (-1)	High Level ($+1$)
A	Prompt Strategy	Direct (rules listed)	CoT (step-by-step preamble)
B	Context	No example provided	One compliant example included
C	Model	Gemini 2.5 Flash-Lite (no thinking)	Gemini 2.5 Pro (with thinking)

1.3 Response Variable

Each run asks the model to answer a fixed content question ("Explain the difference between supervised and unsupervised learning") while obeying six formatting constraints:

Constraint	Rule	Scoring
c1	Do not use the word "important"	Case-insensitive string search
c2	Every sentence ≤ 20 words	Tokenize and count
c3	Exactly 5 bullet points	Count bullet markers
c4	First word is "Interestingly"	Check first token
c5	No question marks	Count "?"
c6	End with "— End of response."	Check last line

Each constraint is scored as 1 (pass) or 0 (fail). The response variable $y = \sum_{i=1}^{6} c_i$ ranges from 0 to 6. All constraints are objectively verifiable, eliminating subjective scoring bias.

1.4 Experimental Controls

- **Randomization**: The 24 runs were executed in a randomized order generated by R (`set.seed(305)`).
- **Independence**: Each run used a fresh API call with no conversation history or context carryover.
- **Fixed content**: The same question was used across all runs to eliminate content difficulty as a confound.
- **Automated scoring**: A Python script scored all six constraints programmatically, ensuring consistency.
- **Temperature**: All API calls used temperature = 1.0 to allow natural variation.

The experiment was executed via the Google Vertex AI REST API, with all responses and metadata (including thinking token counts) recorded for reproducibility.

2 Analysis of the Data

2.1 Factorial Effects

Table 1 presents the estimated factorial effects. The grand mean is 5.67. The pooled MSE from replicates is 0.0833 with 16 degrees of freedom, yielding $SE = 0.1179$ for all effect estimates.

Table 1: Estimated factorial effects with 95% confidence intervals.

	Effect	Estimate	SE	t	p	CI
A	A	0.0000	0.1179	0.0000	1.0000	[-0.250, 0.250]
B	B	0.3333	0.1179	2.8284	0.0121	[0.084, 0.583]
C	C	0.6667	0.1179	5.6569	<0.001	[0.417, 0.916]
AB	AB	0.0000	0.1179	0.0000	1.0000	[-0.250, 0.250]
AC	AC	0.0000	0.1179	0.0000	1.0000	[-0.250, 0.250]
BC	BC	-0.3333	0.1179	-2.8284	0.0121	[-0.583, -0.084]
ABC	ABC	0.0000	0.1179	0.0000	1.0000	[-0.250, 0.250]

Three effects are statistically significant: the main effect of **C (Model)** at 0.6667 ($p < 0.001$), the main effect of **B (Context)** at 0.3333 ($p = 0.0121$), and the **B × C interaction** at -0.3333 ($p = 0.0121$). Factor A (Prompt Strategy) and all its interactions have an estimated effect of exactly zero.

2.2 ANOVA

Table 2: ANOVA table for the 2^3 factorial model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	0.0000	0.0000	0	1.00000
B	1	0.6667	0.6667	8	0.01211
C	1	2.6667	2.6667	32	0.00004
A:B	1	0.0000	0.0000	0	1.00000
A:C	1	0.0000	0.0000	0	1.00000
B:C	1	0.6667	0.6667	8	0.01211
A:B:C	1	0.0000	0.0000	0	1.00000
Residuals	16	1.3333	0.0833	NA	NA

The ANOVA confirms C ($F = 32.0$, $p < 0.001$), B ($F = 8.0$, $p = 0.012$), and B:C ($F = 8.0$, $p = 0.012$) as the only significant sources of variation.

2.3 Plots

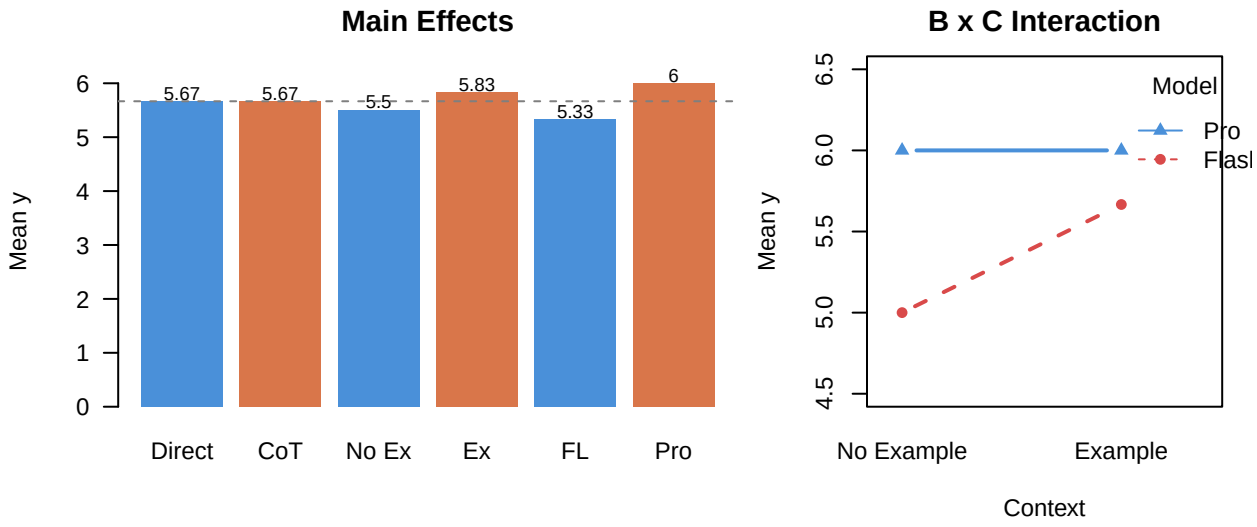

Figure 1: Left: main effects on mean constraints satisfied. Right: B × C interaction—context improves Flash-Lite but not Pro.

The B × C interaction (Figure 1, right) reveals that providing a contextual example improves Flash-Lite's compliance from 5.0 to 5.67, but has no effect on Pro (which remains at 6.0 regardless).

2.4 Diagnostic Assessment

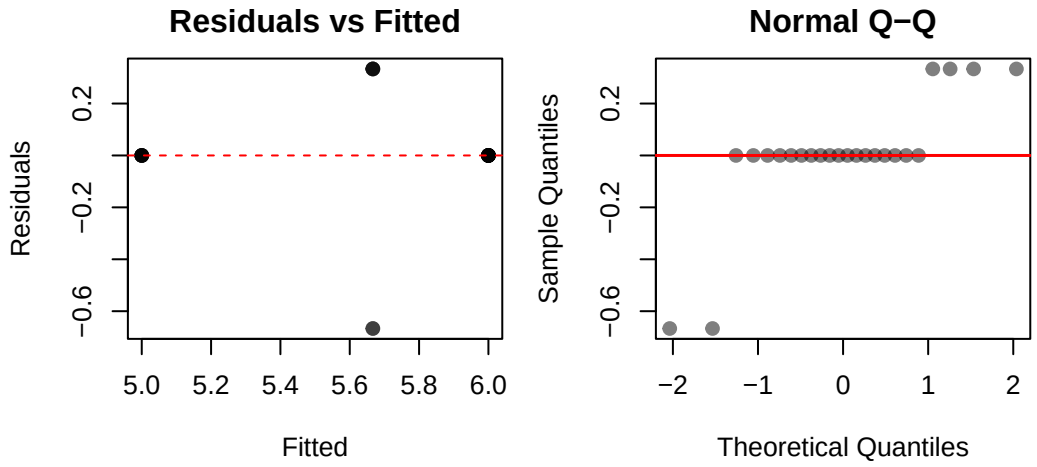

Figure 2: Residual diagnostics. The discrete response and zero-variance groups produce non-standard patterns.

The residual plots show clear departures from normality and constant variance. The response takes only two values (5 or 6), and 6 of 8 treatment groups have zero within-group variance. The pooled MSE (0.0833) is driven entirely by the two Flash-Lite + Example groups. While this violates the standard ANOVA assumptions, the significance of the large effects (C in particular) is robust, as the effect sizes are large relative to any plausible error variance.

2.5 Per-Constraint Analysis

Table 3: Pass rates by constraint and model.

Constraint	Overall	Flash-Lite	Pro
c1: Word ban	1.000	1.000	1
c2: Sentence cap	1.000	1.000	1
c3: 5 bullets	0.667	0.333	1
c4: Opening word	1.000	1.000	1
c5: No questions	1.000	1.000	1
c6: Sign-off	1.000	1.000	1

Constraints c1, c2, c4, c5, and c6 achieved 100% pass rates across all 24 runs. **The only source of variation in the entire experiment is c3** (exactly 5 bullet points). Within Flash-Lite, c3 pass rates are 0/6 without an example and 4/6 with an example, confirming that the contextual example specifically helps the non-thinking model count structural elements correctly.

3 Conclusions

This experiment yields three main findings:

Model reasoning capability (Factor C) is the dominant factor. The Pro model achieved a perfect 6/6 compliance score on all 12 runs, generating an average of 1,725 thinking tokens per run (range: 860–3,312) before producing its response. This internal reasoning—roughly 10–25× more tokens than the visible output—allows the model to parse constraints, draft a candidate, and verify compliance before committing. The Flash-Lite model (zero thinking tokens) failed on at least one constraint in 8 of 12 runs. The effect of C (0.667, $p < 0.001$) is the largest in the experiment, indicating that internal reasoning capability, rather than external prompt engineering, is the primary mechanism enabling reliable instruction-following.

Contextual examples help weaker models (B × C interaction). Providing a compliant example response improved Flash-Lite's performance (5.0 → 5.67) but had no effect on Pro (already at ceiling). The significant B:C interaction (−0.333, $p = 0.012$) indicates that the benefit of few-shot examples is model-dependent: it compensates for limited reasoning in weaker models but is redundant when strong reasoning is available.

Prompt strategy has no effect (Factor A). Whether constraints are presented as a flat list (Direct) or with a step-by-step preamble (CoT) made no measurable difference (effect = 0.000, $p = 1.000$). This is noteworthy because CoT prompting is widely recommended. One interpretation is that when constraints are sufficiently explicit and few in number (six), the model's ability to process them does not depend on how they are formatted in the prompt.

A key limitation is the **ceiling effect** in the Pro group, which compressed all variation into the Flash-Lite conditions and violated the homogeneity of variance assumption. A harder task (more constraints, or more ambiguous constraints) would likely produce greater variance across all groups and enable detection of subtler effects. Despite this, the observed patterns are internally consistent and the significant effects are large enough to be robust to distributional assumptions.