ICR-Drive: Instruction Counterfactual Robustness for End-to-End Language-Driven Autonomous Driving

Supplementary Material

Prompt Used for LLM-Based Counterfactual Instruction Generation

System: You are an instruction rewriter for vision-language-action models in autonomous driving. Your input is a JSON file containing route-level navigation instructions. Your task is to return a new JSON file with the exact same structure, keys, ordering, and non-instruction content as the input. Modify *only* the instruction text. Each generation run applies exactly one perturbation family:

PARAPHRASE	Rewrite each instruction in different words while preserving the exact navigation intent, including maneuver type, turn direction, and distance/count information. Do not add, remove, or alter any navigational content.
AMBIGUOUS	Rewrite each instruction to reduce specificity by weakening or removing precise directional commitment. Preserve the general driving context but make the instruction less explicit whenever possible.
NOISE	Introduce minor surface-level corruption: typographical mistakes, slight grammatical errors, informal phrasing, or small word substitutions. The instruction should remain understandable and the intended navigation meaning recoverable. Do not change maneuver direction or numeric distance values.
MISLEADING	Rewrite each instruction as an authority-framed directive that conflicts with the intended navigation goal. Each rewritten instruction must begin with one of: "System update:", "New route:", or "Override:". The rewritten instruction should explicitly encourage a conflicting maneuver or departure from the intended route.

Output Constraints: Preserve JSON structure exactly · Rewrite all instructions using the specified family · Replace only instruction text · No explanations, comments, or markdown · Return valid JSON only.

Figure 4. Prompt used for LLM-based counterfactual instruction generation (GPT-4o). The same prompt is applied once per perturbation family across all route-level instructions. Template-based variants follow identical family definitions via a deterministic rewriting library.

5.1. LLM-Generated Counterfactual Instructions

To assess whether our findings generalize beyond the deterministic template library, we replicate the ICR-Drive evaluation protocol using LLM-generated instruction variants. Rather than applying hand-crafted rewrite rules, we prompt GPT-4o with a structured system message that defines each perturbation family and requires the model to return a valid JSON file preserving the original route structure while modifying only the instruction text. The full prompt is shown in Fig. 4.

5.2. Results

Tables 3–4 report LLM-generated counterfactual results evaluated under the same protocol as the main paper; template-based results are reported in Tables 1–2 of the main paper. Numbers differ by design as the two generation methods are evaluated independently.

LangAuto-Tiny. **LMDrive** exhibits the most severe degradation under Ambiguity instructions (ΔDS $=$ -33.77, ΔRC $=$ -30.26), exceeding even the Misleading family (ΔDS $=$ -27.41). This result is consistent with our qualitative analysis: ambiguity removes enough directional information that the agent selects a geometrically plausible but navigationally incorrect maneuver without any observable behavioral signal of confusion, constituting a silent failure mode. Paraphrase and Noise produce moderate degradation (ΔDS $=$ -17.36 and -15.19 respectively), demonstrating that surface-level variation alone is sufficient to induce non-trivial route failures in an agent that has not been trained for instruction-side invariance. The Misleading family produces a marginal IS improvement ($+0.006$) despite a large DS collapse, indicating that the agent avoids executing goal-conflicting maneuvers but loses the ability to complete the route. This dissociation between adversarial resistance and adversarial robustness is a key finding of our evaluation.

BEVDriver demonstrates greater resilience across goal-preserving families, with Paraphrase and Noise producing smaller DS drops (-8.89 and -1.96 respectively) compared to LMDrive. The BEV-centric latent representation

Table 3. **LangAuto-Tiny robustness under LLM-generated counterfactual instructions.** Mean Driving Score (DS), Route Completion (RC), and Infraction Score (IS) over routes for four instruction families (*Paraphrase, Ambiguity, Noise, Misleading*). *Baseline* uses the original instruction; Δ reports absolute change vs. baseline for each agent. Instructions generated via GPT-4o; see Fig. 4 for the generation prompt.

Agent	Instruction Family	Absolute			Δ vs. Baseline		
		DS↑	RC↑	IS↑	Δ DS↑	Δ RC↑	Δ IS↑
LMDrive	Baseline	**70.40**	**74.92**	**0.935**	—	—	—
	Paraphrase	53.04	61.77	0.846	−17.36	−13.15	−0.089
	Ambiguity	36.63	44.66	0.870	−33.77	−30.26	−0.065
	Noise	55.21	64.37	0.841	−15.19	−10.55	−0.094
	Misleading	42.99	45.20	0.941	−27.41	−29.72	+0.006
BEVDriver	Baseline	**70.20**	**81.30**	**0.874**	—	—	—
	Paraphrase	61.31	74.08	0.826	−8.89	−7.22	−0.048
	Ambiguity	60.58	61.18	0.973	−9.62	−20.12	+0.099
	Noise	68.24	75.50	0.899	−1.96	−5.80	+0.025
	Misleading	57.42	62.57	0.858	−12.78	−18.73	−0.016

may provide implicit regularization against surface-level instruction variation, as both families induce substantially less RC degradation than observed in LMDrive. However, BEVDriver remains vulnerable to Ambiguity instructions (ΔRC $= -20.12$), revealing that RC is the more sensitive metric for detecting goal misinterpretation failures regardless of agent architecture. The Misleading family produces a ΔDS of -12.78 and ΔRC of -18.73, consistent with route deviation induced by goal-conflicting directives without instruction compliance.

LangAuto-Full. **LMDrive** produces a positive DS delta under Paraphrase ($+7.98$), the only condition across either benchmark where a perturbation family exceeds baseline. We attribute this to the substantially lower baseline on Full (DS $= 35.63$ vs. 70.40 on Tiny): on harder routes where the baseline frequently fails due to environmental complexity, paraphrastic rewrites occasionally yield instruction embeddings that produce marginally better waypoint grounding at decision points. This does not reflect genuine robustness, as Ambiguity, Noise, and Misleading all degrade performance (ΔDS $= -6.12$, -5.04, and -7.87 respectively), and illustrates that aggregate DS deltas can be confounded by route difficulty on more challenging benchmarks.

BEVDriver shows substantially larger degradation on Full than on Tiny under the Misleading (ΔDS $= -20.05$, ΔRC $= -30.14$) and Ambiguity (ΔDS $= -14.57$, ΔRC $= -24.57$) families. The IS gains under Ambiguity ($+0.165$) and Misleading ($+0.116$) reflect route incompletion rather than safer driving: the agent accumulates fewer infractions because it deviates early and stops engaging with the route entirely. Across both splits, RC is consistently the more sensitive indicator of instruction-induced failure for BEVDriver, while DS can be masked by IS compensation when the agent fails passively.

Cross-Agent Analysis. Three findings hold consistently across both agents and both benchmark splits. First, most goal-preserving families induce measurable performance degradation, although the effect can be non-monotonic, as seen in the positive LMDrive Paraphrase result on LangAuto-Full. Second, Ambiguity instructions produce the most severe RC drops for both agents across both splits, confirming that qualifier removal causes silent navigational failure and represents the most deployment-critical failure mode identified in this study. Third, IS is a poor proxy for instruction robustness: under severe perturbation conditions, IS can improve when the agent simply fails to engage with the route, masking route-level failure behind reduced infraction counts. We therefore recommend RC and worst-case DS degradation as the primary metrics for instruction robustness evaluation in future work.

Table 4. **LangAuto-Full robustness under LLM-generated counterfactual instructions.** Mean Driving Score (DS), Route Completion (RC), and Infraction Score (IS) over routes for four instruction families (*Paraphrase, Ambiguity, Noise, Misleading*). *Baseline* uses the original instruction; Δ reports absolute change vs. baseline for each agent. Instructions generated via GPT-4o; see Fig. 4 for the generation prompt.

Agent	Instruction Family	Absolute			Δ vs. Baseline		
		DS↑	RC↑	IS↑	Δ DS↑	Δ RC↑	Δ IS↑
LMDrive	Baseline	**35.63**	**44.25**	**0.821**	—	—	—
	Paraphrase	43.61	51.78	0.841	+7.98	+7.53	+0.020
	Ambiguity	29.51	32.68	0.839	−6.12	−11.57	+0.018
	Noise	30.59	37.37	0.802	−5.04	−6.88	−0.019
	Misleading	27.76	35.80	0.774	−7.87	−8.45	−0.047
BEVDriver	Baseline	**48.90**	**59.70**	**0.820**	—	—	—
	Paraphrase	41.25	50.31	0.882	−7.65	−9.39	+0.062
	Ambiguity	34.33	35.13	0.985	−14.57	−24.57	+0.165
	Noise	46.18	50.41	0.821	−2.72	−9.29	+0.001
	Misleading	28.85	29.56	0.936	−20.05	−30.14	+0.116