Over-Generation and Compaction: A Prompting Strategy for Procedural Text Adaptation with Large Language Models

Hyeongsik Kim, Yanheng Xu, Chaoqun Dong, Fei Du

Bosch Corporate Research and BSH Hausgeräte GmbH

EMNLP 2025 Suzhou, China 中国苏州

1 Introduction

Motivation

1. **Goal:** Procedural text adaptation (recipes, repair guides) that propagates constraints (ingredient/part swaps, timing, safety) while preserving structure and feasibility
2. **Pain points:** Single-pass LLM prompts often produce superficial edits (vestigial steps, wrong tools/seasonings, order violations). Alignment-induced brevity and style biases suppress latent domain knowledge. Fine-tuned, domain-specific systems rely on curated resources and do not scale across domains or variations
3. **Our stance:** Use general-purpose LLMs with prompting scaffolds that (i) expand the search space, (ii) enforce target-format fidelity, and (iii) require no further fine-tuning

Known Approach

1. Graph/knowledge-graph substitutions; hierarchical editing; user-guided critiquing; task-specific fine-tuning, ...
   - Limitations: dependence on curated resources, narrow coverage, brittle out-of-domain generalization
2. LLM prompting & scaffolding:
   - Single-step rewrite and style transfer: fast but prone to vestigial steps and shallow edits
   - Chain-of-Thought (CoT): exposes reasoning but often fails to translate plans into faithful procedural rewrites
   - Self-critique / Self-Refine loops: iterative quality gains; risk of conservatism and residual errors; higher cost; unclear stopping

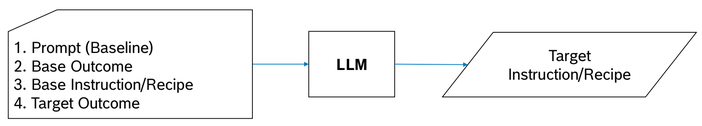
1. Prompt (Baseline)

You are a professional content creator. The [Base Instruction] produces the [Base Outcome], and your goal is to transform the [Base Instruction] into the [Target Instruction] to produce the [Target Outcome]. Output only the final transformed [Target Instruction] without any additional commentary or explanations.

Input Parameters:
- [Base Outcome]: “Steamed Turbot (清蒸多宝鱼)”
- [Base Instruction]: “Gather all the ingredients. ... Place the fish on the steaming rack, and steam for about 8 minutes. Meanwhile, in an Amway Queen 1.4L pan, heat a suitable amount of Amway Queen extra virgin olive oil, add chopped scallions and the seasoned soy sauce for seafood, and stir-fry until fragrant. Once the fish is done steaming, sprinkle some shredded scallions, goji berries, and cilantro. Finally, drizzle the hot fragrant oil over the top! Perfect!”
- [Target Outcome]: “Steamed Hairy Crab (清蒸大闸蟹)”

LLM

[Target Instruction]: Gather all the ingredients. ... Steam the crab for about 15-20 minutes, depending on its size. Meanwhile, in an Amway Queen 1.4L pan, heat a suitable amount of Amway Queen extra virgin olive oil, add chopped scallions, ginger, and a touch of seasoned soy sauce for seafood, and stir-fry until fragrant. Once the crab is done steaming, garnish with shredded scallions and cilantro. Finally, drizzle the hot fragrant oil over the top! Perfect!



2 Solution

1. Inputs: base instruction (e.g., source recipe), base outcome (e.g., source dish), target outcome (e.g., target dish)
2. Stage 1 (Over-generation): Unconstrained, verbose intermediate to surface target-specific procedural knowledge (e.g., tools, timings, checks, edge cases)
3. Stage 2 (Compaction): Condense into Target Instruction that matches the length/format/tone of base instruction while preserving target constraints
4. Small-model pathway (2-step compaction): Distillation and then style-transfer compaction

1. Decouple knowledge elicitation from style/format constraints. First expand to surface target-specific procedural knowledge; then compress to match the base format
2. Over-generation counters alignment/brevity biases, reduces vestigial steps, and exposes edge cases and safety checks. Compaction enforces fidelity to base length, tone, and structure while propagating target constraints
3. Separation improves edit fidelity and robustness across model scales; small models benefit from explicit selection-before-style transfer

1. Prompt: Over-generation

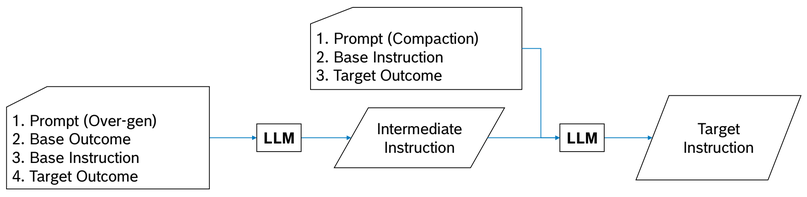
You are a professional content creator. The [Base Instruction] produces the [Base Outcome], and your goal is to transform the [Base Instruction] into the [Intermediate Instruction] to produce the [Target Outcome]. Do not worry about the format, structure, or length of the [Intermediate Instruction]. However, your output must be a comprehensive, expanded version of the [Intermediate Instruction] that is at least **1000** times longer than the original [Base Instruction]. You must incorporate all relevant knowledge, information, best practices, important considerations, as well as warnings or pitfalls to avoid — essentially, everything you know about the subject, Instruction, and outcome. Output only the final [Intermediate Instruction], with no additional commentary or explanation. ...

2. Prompt: Compaction

You are a professional content creator working with the [Intermediate Instruction], which contains rich details intended to achieve the [Target Outcome]. Your top priority is to strictly match the length and format of the [Base Instruction], disregarding the original format of the [Intermediate Instruction]; within these constraints, condense the [Intermediate Instruction] as effectively as possible into the [Target Instruction]. Provide only the condensed [Target Instruction] without commentary, introduction, or explanation.

LLM

[Target Instruction]: ... Tie their legs and claws securely with kitchen twine to prevent movement during steaming. Fill a wok with water, add sliced ginger to the water, and bring it to a boil. Place the tied crabs belly-up on a steaming rack, ensuring they are spaced apart for even cooking. Cover the wok with a tight-fitting lid and steam over medium heat for 15-20 minutes, depending on the size of the crabs. Meanwhile, prepare the dipping sauce by combining light soy sauce, black vinegar, a pinch of sugar, sesame oil, and finely chopped ginger and scallions. Once the crabs are done steaming, their shells will turn bright orange, and their aroma will be fragrant. Transfer the crabs to a serving plate, garnish with scallions and cilantro, and serve with the dipping sauce on the side. Perfect!



3 Evaluation

		ChatGPT 4o				DeepSeek V3			
Group	LLM	Mean	Med	Std	95% CI	Mean	Med	Std	95% CI
Source		7.16	8	1.93	[7.08, 7.23]	5.67	6	1.93	[5.59, 5.74]
Base	D1	8.40	9	1.36	[8.34, 8.45]	5.85	5	1.99	[5.77, 5.93]
	4o	7.30	8	1.87	[7.22, 7.37]	4.66	5	1.49	[4.60, 4.72]
	4o-mini	6.93	7	1.94	[6.86, 7.01]	4.54	4	1.43	[4.49, 4.60]
	M	6.90	7	1.89	[6.83, 6.98]	4.43	4	1.40	[4.37, 4.48]
	D2	6.24	7	1.70	[6.17, 6.31]	2.17	2	1.27	[2.12, 2.22]
CoT	4o	7.62	8	1.53	[7.57, 7.69]	4.41	5	2.18	[4.33, 4.50]
	4o-mini	7.44	8	1.43	[7.38, 7.49]	4.08	5	2.29	[3.99, 4.17]
	M	6.67	7	1.73	[6.60, 6.73]	3.40	4	2.01	[3.32, 3.48]
Critics	**4o**	**8.81**	**9**	**0.88**	**[8.78, 8.84]**	6.76	7	2.01	[6.69, 6.84]
	4o-mini	8.59	9	0.87	[8.56, 8.63]	6.48	7	2.02	[6.40, 6.56]
	M	8.24	9	1.23	[8.20, 8.29]	6.17	6	2.04	[6.09, 6.25]
OC	4o	8.74	9	1.02	[8.70, 8.78]	6.83	7	2.02	[6.74, 6.92]
	4o-mini	**8.80**	**9**	**1.12**	**[8.76, 8.84]**	**8.03**	**9**	**2.12**	**[7.94, 8.13]**
	M	7.62	9	2.20	[7.53, 7.70]	7.43	8	1.91	[7.35, 7.52]

Table 1: Recipe adaptation quality under the RCF metric (higher is better). Results are reported side-by-side for ChatGPT-4o and DeepSeek v3 evaluations across prompting strategies. ’Base’ refers to baseline methods; ’4o’ and ’4o-mini’ denote ChatGPT-4o and 4o-mini; ’M’ stands for Mistral-7B; ’D1’ and ’D2’ refer to DeepSeek-r1 671B and 7B, respectively

Group	LLM	Mean	Med	Std	95% CI	Mean	Med	Std	95% CI
Source		7.03	7	1.59	[6.89, 7.17]	6.57	7	1.81	[6.41, 6.72]
Base	D1	8.05	8	1.22	[7.95, 8.16]	7.20	8	1.73	[7.05, 7.36]
	4o	7.30	8	1.58	[7.16, 7.44]	6.62	7	1.79	[6.46, 6.78]
	4o-mini	7.09	7	1.61	[6.95, 7.23]	6.51	6	1.77	[6.35, 6.66]
	M	6.86	7	1.54	[6.73, 7.00]	6.14	6	1.66	[5.99, 6.28]
	D2	4.93	6	1.60	[4.79, 5.07]	3.62	4	1.29	[3.51, 3.74]
CoT	4o	7.31	8	1.43	[7.18, 7.44]	5.93	6	2.19	[5.73, 6.12]
	4o-mini	7.11	7	1.36	[6.99, 7.23]	5.91	6	2.25	[5.72, 6.11]
	M	6.38	6	1.63	[6.23, 6.52]	5.38	5	2.00	[5.21, 5.56]
Critics	4o	8.42	9	1.09	[8.32, 8.51]	7.95	8	1.47	[7.82, 8.08]
	4o-mini	8.53	9	0.80	[8.46, 8.60]	8.12	8	1.38	[8.00, 8.24]
	M	7.43	8	1.49	[7.30, 7.57]	6.70	7	2.18	[6.51, 6.89]
OC	4o	8.37	9	1.23	[8.26, 8.48]	7.77	8	1.64	[7.63, 7.91]
	4o-mini	**8.70**	**9**	**1.03**	**[8.61, 8.79]**	**8.82**	**9**	**1.46**	**[8.69, 8.95]**
	M	7.91	9	1.72	[7.76, 8.06]	7.85	8	1.74	[7.70, 8.01]

Table 2: myfixit fixing-instruction quality under the adapted RCF-style metric (higher is better)

Metric details (RCF: Recipe Consistency & Feasibility)

Score procedural validity and executability of adapted instructions; prioritize feasibility over lexical similarity (Scale: Integer 1–10 (higher is better). 1 = unsafe/incoherent; 10 = clear, complete, executable)

Evaluator protocol:
- LLM-as-judge using g-eval-style prompts with Chain-of-Thought and fixed rubrics
- Dual evaluators to reduce bias: ChatGPT-4o and DeepSeek-v3; identical instructions and scoring criteria

Rubric dimensions (assessed sequentially):
- Ingredient/part–action consistency: all listed items used; no phantom ingredients
- Irrelevant/unsuitable actions: remove physically impossible or context-inappropriate steps
- Essential prep & safety: include critical prep, safety checks, readiness/verification cues
- Logical step order: maintain causal dependencies (prep → cook → finish)
- Harmful instructions: flag unsafe temps/tools/cross-contamination risks
- Brevity vs. completeness: concise but not at the expense of success-critical details

Key Takeaways

- Two-stage prompting beats single-pass for procedural edits: expand first to unlock latent knowledge, then compact to enforce target constraints
- OC reliably removes vestigial steps and inserts target-specific checks (tools, timings, safety) while matching base style and length
- RCF prioritizes feasibility over fluency; report results with dual evaluators to mitigate judge calibration differences
- 4o-mini + OC offers a strong accuracy–cost balance; for small models, use two-step compaction (select essentials → style transfer)