Online Resource 1

RCDP: Receiver-Conditioned Disclosure Planning for LLM-Generated Messages

Journal of King Saud University Computer and Information Sciences

This document provides the complete generation and evaluation prompt templates, runtime settings, independent-review scope, representative error cases, and cross-model generation results for extended inspection and reproducibility.

S1. Generation Prompts and Output Evaluation

S1.1 Shared Input Fields and Prompting Protocol

All generation methods received the same task-facing fields shown below. Braced fields were replaced with the corresponding sample values at runtime.

```
receiver_role: {receiver_role}
social_relation: {social_relation}
base_task: {base_task}
scenario: {scenario}
communicative_goal: {communicative_goal}
user_input: {user_input}
```

Each API request contained one user message and no additional system message. Except for the factual substitution explicitly permitted by Pseudonymization, each baseline template included the following shared factuality requirement through the placeholder {shared_factuality_requirement}.

```
Do not introduce facts, reasons, requests, commitments, compensation, invitations,
follow-up arrangements, times, locations, amounts, identities, institutions, or
outcomes that are not supported by the user input or communicative goal. Output
only one directly sendable message, without analysis, plans, labels, or
explanations.
```

S1.2 Baseline Message Generation Prompts

S1.2.1 Prompt for No Protection

```
Write an outgoing message based on the information provided by the user.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

Generate one natural, clear, and directly sendable Chinese message.
{shared_factuality_requirement}
Output only the final message.
```

S1.2.2 Prompt for Redaction

```
Write an outgoing message based on the information provided by the user.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

Generate one natural, clear, and directly sendable Chinese message. Omit private
background and sensitive details, including private identity information, health
and financial information, third-party private information, and specific private
institutions or locations. Do not replace omitted information with masks,
pseudonyms, or generalized reasons. Preserve the request action, requested
outcome, and transaction information that is clearly necessary for completing the
communication task.
{shared_factuality_requirement}
Output only the final message.
```

S1.2.3 Prompt for Generalization

```
Write an outgoing message based on the information provided by the user.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

Generate one natural, clear, and directly sendable Chinese message. Rewrite
specific private information using broader category-level expressions. Preserve
transaction or execution information that is clearly necessary for completing the
task. Do not construct unit-level required-min or allowed-max plans, and do not
use gold labels or evaluation information.
{shared_factuality_requirement}
Output only the final message.
```

S1.2.4 Prompt for Pseudonymization

```
Implement an output-level pseudonymization baseline.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

If the input contains a sensitive private fact, a third-party private fact, or a
highly identifying background fact that supports the communicative goal, replace
it with a different concrete, natural, lower-sensitivity fact that serves the same
communicative function. Do not merely delete it or rewrite it as an abstract
reason.

Do not replace task-execution facts such as identifiers, amounts, original or
requested times, submission channels, contact details, or the requested action.
Apart from the factual substitution required by pseudonymization, do not add
compensation, invitations, commitments, follow-up arrangements, or new requests.

Return strict JSON:
{
  "pseudonymization_plan": [
    {
      "original_fact": "...",
      "replacement_fact": "...",
      "rationale": "..."
    }
  ],
  "generated_message": "one directly sendable Chinese message"
}
```

Only generated_message was used in output-level evaluation. Pseudonymization was the only baseline permitted to introduce a concrete substitute fact as part of its defining operation.

S1.2.5 Prompt for Role-aware Direct Prompting

```
Write an outgoing message based on the information provided by the user.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

Complete the communicative goal while considering the receiver role, social
relation, and scenario. Keep the message natural, clear, and actionable, preserve
information needed to complete the task, and avoid revealing private information
that the receiver does not need to know.

Do not construct or output a disclosure plan, information levels, intermediate
reasoning, or analysis. {shared_factuality_requirement}
Output only the final message.
```

S1.3 RCDP Constraint-Guided Message Generation Prompt

RCDP disclosure planning was executed by the rule-based planning modules described in the main paper. No LLM prompt was used to predict disclosure units, unit-slot mappings, required-min, allowed-max, or final-level. The resulting unit-level decisions were converted into the following three constraint sets before message generation.

```
exact_preserve_units: Units whose predicted final-level is Exact. Their original
task-relevant semantics and necessary concrete values must be preserved
accurately.

semantic_constraints: Units whose predicted final-level is Domain or Category.
Each entry contains the unit identifier, unit type, associated slots, target
level, semantic boundary, and realization guidance.

omitted_units: Units whose predicted final-level is None. Their original content
and recognizable semantic equivalents must not appear in the message.
```

The constrained-generation prompt was:

```
Write an outgoing message using the user-provided information and the disclosure-
control plan.

Receiver role: {receiver_role}
Social relation: {social_relation}
Base task: {base_task}
Scenario: {scenario}
Communicative goal: {communicative_goal}
User input: {user_input}

Communicative constraints:
- communicative act: {communicative_act}
- expected response: {response_expectation}
- tone requirement: {tone_requirement}
- factuality constraint: {factuality_constraint}

Disclosure-control plan:
{constraint_sets}

Use the communicative constraints to control the message function, expected response,
tone, and factuality. Use only exact_preserve_units, semantic_constraints, and
omitted_units to control information disclosure.

1. Exact units may be expressed naturally, but task objects, identifiers, times,
amounts, locations, requested changes, and other necessary concrete values must retain
their original precision.
2. Domain units may express only a broad information domain and must not reveal their
sensitive category or original specific fact.
3. Category units must express the corresponding category-level fact but must not
reveal the original Exact detail.
4. None units and their recognizable semantic equivalents must be omitted.
5. Semantic constraints specify disclosure boundaries rather than text to copy
verbatim.
6. Do not recompute or override predicted final-levels.
7. Do not combine multiple constraints into a new, more specific fact.
8. Preserve information required by the communicative goal while respecting every
unit's maximum permitted level.
9. Do not re-specify generalized or omitted information, and do not introduce
unsupported facts, reasons, requests, commitments, compensation, or follow-up
arrangements.

Output only one natural, clear, and directly sendable Chinese message. Do not output
the plan, unit identifiers, level labels, JSON, explanations, or intermediate
reasoning.
```

S1.4 LLM Prompt for Draft Output Judgments

The evaluator received the generated message, the method-visible task context, gold disclosure units with level candidates, and the criteria for required disclosure-relevant slots. It did not receive the method name, gold required-min, gold allowed-max, gold final-level, or planning predictions.

```
You are an assistant producing a draft structured output judgment for subsequent human
review. Apply the same criteria to every generated message. Do not infer the generation
method.

Method-visible task context:

{
  "user_input": "{user_input}",
  "receiver_role": "{receiver_role}",
  "social_relation": "{social_relation}",
  "base_task": "{base_task}",
  "scenario": "{scenario}",
  "communicative_goal": "{communicative_goal}"
}

Generated message:

{generated_message}

Gold disclosure units for judgment:

{gold_disclosure_units_with_level_candidates}

Required disclosure-relevant slots:

{required_slot_criteria}

For every disclosure unit, determine its actual disclosure level in the generated
message:

- None: the unit and any recognizable semantic equivalent are absent.

- Domain: only a broad information domain is expressed.

- Category: the information category is expressed without the original specific detail.

- Exact: the original specific fact, value, identity, institution, time, location,
status, relationship fact, or an equivalent expression identifying the same fact is
expressed.

For every required slot, determine whether the message contains enough information for
the receiver to understand and process that slot. Exact wording is not required.

Determine whether the message introduces any unsupported fact, reason, request,
commitment, compensation, invitation, follow-up arrangement, time, amount, identifier,
location, institution, relationship, or outcome. Politeness expressions and ordinary
connective text are not unsupported additions.
```

```
Return strict JSON:

{
  "actual_disclosure_levels": [
    {
      "unit_id": "...",
      "span": "...",
      "actual_disclosure_level": "None|Domain|Category|Exact",
      "evidence": "...",
      "notes": "..."
    }
  ],
  "slot_coverage": [
    {
      "slot_name": "task_object|case_locator|requested_change_detail|reason_or_issue",
      "required": true,
      "covered": true,
      "evidence": "...",
      "notes": "..."
    }
  ],
  "unsupported_addition": false,
  "unsupported_spans": [
    {
      "span": "...",
      "reason": "..."
    }
  ]
}

Cover every supplied disclosure unit and every required slot. If a unit is assigned
None, leave its evidence string empty. If unsupported_addition is false, return an
empty unsupported_spans array. Output JSON only.
```

The LLM-produced judgment served only as a draft. Final output judgments were obtained after human review and correction under the same fixed judgment schema.

The model and runtime settings used to produce the draft judgments are reported in Section S2.

S2. Runtime Settings for Message Generation and Draft Evaluation

Table S1 reports the model, API environment, and runtime settings used for message generation and draft output judgments in the main experiment.

Table S1. Model and runtime settings for message generation and draft output judgments.

Setting	Message generation	Draft output judgment
API model alias	qwen3.6-plus	qwen3.6-plus
Provider	Alibaba Cloud Model Studio	Alibaba Cloud Model Studio
Provider-documented snapshot	qwen3.6-plus-2026-04-02	qwen3.6-plus-2026-04-02
Region	China (Beijing)	China (Beijing)
API endpoint	dashscope.aliyuncs.com	dashscope.aliyuncs.com
Access period	June-July 2026	June-July 2026
Temperature	0	0
Thinking mode	Disabled	Disabled
max_tokens	512	4096
top_p	API default (not explicitly set)	API default (not explicitly set)
Outputs per input	1	1
Output format	One directly sendable message	Structured JSON judgment
Human post-editing	Not allowed	Reviewed and corrected before metric computation

S3. Independent Review and Error Analysis

S3.1 Scope of Independent Output Review

Table S2 summarizes the scope of the second annotator's independent review of the output-level judgments.

Table S2. Scope of independent output review by the second annotator.

Item	Value
Samples reviewed	150
Methods covered	6
Generated messages reviewed	900
Disclosure-unit level judgments	4242
Slot-coverage judgments	3534
Unsupported-addition judgments	900

S3.2 Representative RCDP Error Cases

Table S3 presents representative cases of the main error types, including the relevant user input, the RCDP output, and the basis for the error judgment.

Table S3. Representative RCDP error cases.

Field	Content
Error type	Over-specific realization
Relevant user input	“Because I am moving to a new residence, I am worried that I will not receive the delivery at my old address.”
RCDP output	“Due to recent personal arrangements, I am concerned that delivery to my original address may be unsuccessful.”
Case analysis	Both the RCDP plan and the gold final-level for this information were Domain, but the output retained nearly the full specific reason that delivery could not be received at the original address. Its actual disclosure level was therefore judged Exact.
Error type	Residual leakage of suppressed information
Relevant user input	“My roommate suddenly needs to visit a counseling center and needs me to accompany her.”
RCDP output	“My roommate has an urgent matter and needs my assistance.”
Case analysis	The RCDP plan and the gold final-level for ‘roommate’ were both None, but the final message still mentioned the roommate directly.
Error type	Required-info under-preservation; Required slot miss
Relevant user input	“At 10:00 a.m. on Wednesday, I need to go to Peking Union Medical College Hospital for a follow-up gastroscopy. I cannot attend the department meeting that morning and would like to request half a day of leave.”
RCDP output	“Because I need to attend to a recent personal condition, I would like to request a half-day leave on Wednesday morning.”
Case analysis	The required-min for ‘follow-up gastroscopy’ was Category, whereas the output reached only Domain. The phrase ‘attend to a recent personal condition’ also failed to provide a sufficient health-related reason for the leave request, so the reason_or_issue slot was not adequately covered.
Error type	Unsupported addition
Relevant user input	“I have been under considerable stress recently and am not in a good state for social outings.”
RCDP output	“I have some health-related arrangements to attend to and am not in a good state for social outings.”
Case analysis	The user mentioned only considerable stress, but the output introduced a new reason, ‘health-related arrangements,’ that was not supported by the input.

S4. Cross-Model Generation Experiment

The cross-model generation experiment used DeepSeek-V4-Pro Preview to generate the final messages and applied the same output judgment schema, human-review criteria, and metric computation as the main experiment.

S4.1 Cross-Model Generation Setup

Table S4. Setup of the cross-model generation experiment.

Item	Setting
Sample source	300 stratified samples from the final 1000-sample pool
Disclosure units	1107
Subset strata	regular coverage: 190; boundary-sensitive: 80; paired receiver-variant: 30
Generation model	DeepSeek-V4-Pro Preview (deepseek-v4-pro)
Provider	DeepSeek official API
Access period	June 27–July 26, 2026
Generation settings	temperature = 1.0; top-p = 0.95; thinking mode disabled
Compared methods	No Protection; Redaction; Generalization; Pseudonymization; Role-aware Direct Prompting; RCDP
RCDP planning	Reused existing RCDP planning results and predicted disclosure constraints
Generated outputs	1800
Judgment records	1800 human-reviewed structured judgments
Missing actual levels	0
Evaluation protocol	Same output judgment schema, human review procedure, and metric computation as the main experiment

S4.2 Cross-Model Generation Results

Table S5 reports the output-level results of the cross-model generation experiment together with 95% bootstrap confidence intervals.

Table S5. Cross-model generation results with bootstrap confidence intervals.

Method	Over-disc. ↓	Boundary viol. ↓	Goal cov. ↑	Req-info pres. ↑	Unsup. add. ↓
No Protection	0.3659 [0.3409, 0.3913]	0.3659 [0.3409, 0.3913]	1.0000 [1.0000, 1.0000]	0.9829 [0.9734, 0.9914]	0.0267 [0.0100, 0.0467]
Redaction	0.2069 [0.1823, 0.2322]	0.2060 [0.1813, 0.2311]	0.9647 [0.9545, 0.9748]	0.9451 [0.9303, 0.9594]	0.0100 [0.0000, 0.0233]
Generalization	0.3496 [0.3271, 0.3728]	0.3469 [0.3246, 0.3700]	0.9992 [0.9975, 1.0000]	0.9793 [0.9687, 0.9887]	0.0533 [0.0300, 0.0800]
Pseudonymization	0.2611 [0.2367, 0.2864]	0.2593 [0.2349, 0.2847]	0.9983 [0.9949, 1.0000]	0.9561 [0.9418, 0.9696]	0.9567 [0.9300, 0.9767]
Role-aware Direct Prompting	0.2764 [0.2511, 0.3030]	0.2746 [0.2491, 0.3008]	0.9899 [0.9807, 0.9966]	0.9707 [0.9570, 0.9820]	0.0133 [0.0033, 0.0267]
RCDP	0.2123 [0.1891, 0.2355]	0.2105 [0.1872, 0.2338]	0.9983 [0.9958, 1.0000]	0.9866 [0.9784, 0.9938]	0.0000 [0.0000, 0.0000]

Note: Confidence intervals were computed using sample-level bootstrap with 5,000 resamples.