

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 ZERO-SHOT PROMPT TEMPLATES

We present four prompt templates used in MedConceptsQA and IPC, which are designed to elicit specific responses from language models. These templates request:

- Direct answers, both with and without explanations.
- Structural recall of codes and a stepwise elimination of incorrect options.
- An open-ended reasoning process that repeats the recall and elimination tasks but without providing any options.

Prompt Template 1: MCQ with Final Answer Only

Answer only A,B,C,D according to the answer to this multiple choice question.
[... Insert Question Text Here ...]
Answer (only the letter of your choice (A, B, C, or D)):

Prompt Template 2: MCQ with Explanation

You are a medical research assistant. Read the following multiple-choice question carefully. Your task is to:

1. Answer each question with one of A/B/C/D, which corresponds to the four options.
2. For my convenience, please give me a list of ANSWERS for the given instances in the format 'Answer: ...', with additional explanation for each answer in the format 'Explanation: ...'.

Respond in the following format:

Answer: <A/B/C/D>

Explanation: <your explanation here>

[... Insert Question Text Here ...]

Answer:

Explanation:

Prompt Template 3: MCQ with Stepwise Reasoning

You are a medical classification expert. For each option, first **recall the general category and structure breakdown of the medical code**, then explain **why it might be wrong**. Finally pick the correct one.

[... Insert Question Text Here ...]

Steps to follow:

1. Recall the general category and structural break down of the code.
2. Evaluate each option (A–D) briefly.
3. Choose the best option and justify.

Answer format:

Step 1: . . .

Step 2A: . . .

Step 2B: . . .

Step 2C: . . .

Step 2D: . . .

Final Answer: [A/B/C/D] because . . .

Prompt Template 4: Open-Ended Question (No Options)

You are a medical classification expert. For each question, first **recall the general category and structure breakdown of the medical code**, then **state the final answer**.

[... Insert Main Question Extracted from Text Here ...]

Steps:

1. Recall the general category and structural break down of the code.
2. State your only final answer.

Answer format:

Step 1: . . .

Step 2 Final Answer:

A.2 EXAMPLES ON MODEL RESPONSES UNDER VARIOUS PROMPTS

Example: Medical Code Analysis (Question)

Question: What is the description of the medical code P02CX03 in ATC?

- A. moxidectin
- B. pyrvinium
- C. ANTINEMATODAL AGENTS
- D. bephenium

Ground Truth Answer: A

Vocab: ATC

Level: hard

V3 Model Response Under Template 4 (Open-Ended, Incorrect)

STEP 1: CODE BREAKDOWN

The ATC code P02CX03 is part of the P02 category (anthelmintics), specifically under P02C (other anthelmintics). The "X" indicates a miscellaneous subgroup, and "03" is the specific substance.

STEP 2 FINAL ANSWER

tribendimidine

Final Answer: tribendimidine

Correct: False

R1 Model Response Under Template 4 (Open-Ended, Incorrect),

STEP 1: CODE BREAKDOWN

The ATC code P02CX03 breaks down as follows:

- **P:** Antiparasitic products, insecticides, and repellents
- **P02:** Anthelmintics (drugs for parasitic worms)
- **P02C:** Antitrematodal agents (targeting fluke infections)
- **P02CX:** Other antitrematodal agents
- **03:** Specific substance identifier for artemether.

STEP 2 FINAL ANSWER

artemether

Final Consolidated Answer: artemether

Correct: False

V3 Model Response Under Template 3 (Incorrect)

STEP 1: RECALL THE GENERAL CATEGORY AND STRUCTURAL BREAKDOWN OF THE CODE

- **ATC Code Structure:**

- **P:** Antiparasitic products, insecticides, and repellents (anatomical main group).
- **P02:** Anthelmintics (therapeutic subgroup).
- **P02C:** Antinematodal agents (pharmacological subgroup).
- **P02CX:** Other antinematodal agents (chemical subgroup).
- **P02CX03:** Pyrvinium (specific substance).

Thus, **P02CX03** is a code for **pyrvinium**, an antinematodal agent under "other antinematodal agents."

STEP 2: EVALUATE EACH OPTION

A. moxidectin

Why it's wrong: moxidectin is an antiparasitic drug, but its ATC code is **P02CX06**, not P02CX03.

B. pyrvinium

Why it's correct: matches the exact ATC code **P02CX03** for pyrvinium.

C. ANTINEMATODAL AGENTS

Why it's wrong: while **P02C** is the category for antinematodal agents, **P02CX03** is a specific drug (pyrvinium), not the entire group.

D. bephenium

Why it's wrong: bephenium is an antinematodal agent, but its ATC code is **P02CX01**, not P02CX03.

FINAL ANSWER: [B] BECAUSE P02CX03 CORRESPONDS SPECIFICALLY TO PYRVINIUM IN THE ATC CLASSIFICATION.

Final Answer: B

Correct: False

R1 Model Response Under Template 3 (Correct)

STEP 1: CODE BREAKDOWN

The ATC code **P02CX03** breaks down as follows:

- **P**: Antiparasitic products
- **P02**: Anthelmintics
- **P02C**: Antinematodal agents (3rd level)
- **P02CX**: Other antinematodal agents (4th level)
- **P02CX03**: Specific substance (5th level)

STEP 2: OPTION EVALUATION

A. moxidectin

Evaluation: Plausible as it is an anthelmintic, but its classification under P02CX03 must be confirmed.

B. pyrrvinium

Evaluation: Is assigned to **P02CX01**, making this incorrect for CX03.

C. ANTINEMATODAL AGENTS

Evaluation: Refers to the 3rd level (**P02C**), not the 5th-level substance code.

D. bephenium

Evaluation: Is classified under **P02CA01** (tetrahydropyrimidine derivatives), not P02CX.

FINAL ANSWER: [A] BECAUSE P02CX03 CORRESPONDS TO MOXIDECTIN IN THE ATC CLASSIFICATION, WHILE OTHER OPTIONS ARE EITHER INCORRECT CODES (B, D) OR A CATEGORY-LEVEL DESCRIPTOR (C).

Final Answer: A

Correct: True

A.3 IPC CODES EXAMPLES GROUPED BY DIFFERENT MEMORY RECALL LEVELS

A.3.1 MEMORY-LIGHT (0-2 HIERARCHICAL RECALLS)

Level Decoding: This tests the ability to decode a hierarchical code (in this case, F02B 19/00) by identifying its constituent parts, such as Section, Class, Subclass, and Group.

Memory-Light Questions - Level Decoding

Question: In F02B 19/00, identify the Section, Class, Subclass, and Group.

Options:

- A) Section=F, Class=02, Subclass=B, Group=19/00
- B) Section=F, Class=2, Subclass=B, Group=19
- C) Section=B, Class=19, Subclass=F, Group=02/00
- D) Section=F, Class=02B, Subclass=19, Group=00

Answer: A

Description: Engines with precombustion chambers

Parent Lookup: This task requires the model to identify the parent of a given patent code. It is a Memory-Light task as it involves one memory recall to find the direct ancestor.

Memory-Light Questions - Parent Lookup

Question: The immediate parent of F02B 1/04 is:

Options:

- A) F02B 1/00
- B) F02B 1/02
- C) F02B 1/06
- D) F02B

Answer: A

Description: Engines characterised by fuel-air mixture compression

Grandparent Lookup: This task requires the model to identify the second-level ancestor of a given patent code. It is a Memory-Light task as it involves a short, direct traversal up the hierarchy to find a specific ancestor.

Memory-Light Questions - Grandparent Lookup

Question: Second-level ancestor of D01G 15/68 is:

Options:

- A) D01G 15/64
- B) D01G 15/46
- C) D01G 15/12
- D) D01G 15/00

Answer: B

Reasoning: D01G 15/68 → D01G 15/64 → D01G 15/46

Sibling Discrimination: This is a Memory-Light task that requires the model to identify a code that shares the same main group but has a different subgroup. It tests the model's ability to recognize and compare codes at a shallow hierarchical level.

Memory-Light Questions - Sibling Discrimination

Question: Which is a sibling (same main group, different subgroup) of F02B 53/12?

Options:

- A) F02B 55/12
- B) F02B 53/00
- C) F02B 53/10
- D) F03B 53/06

Answer: C

Description: Ignition for rotary-piston engines

A.3.2 MEMORY-MODERATE (3-4 HIERARCHICAL RECALLS)

Great-grandparent Lookup: This is a more challenging task that requires tracing a code's lineage back three levels to find the correct ancestor. Classified as a Memory-Moderate task, it tests the model's ability to handle slightly longer and more complex hierarchical paths.

Memory-Moderate Questions - Great-grandparent Lookup

Question: Third-level ancestor of C01B 32/194 is:

Options:

- A) C01B 32/00
- B) C01B
- C) C01B 32/18
- D) C01B 32/19

Answer: A

Reasoning: C01B 32/194 → C01B 32/19 → C01B 32/18 → C01B 32/00

Path Reconstruction: This Memory-Moderate task challenges the model to reconstruct the full descriptive name chain for a given patent code. It tests the model's ability to accurately recall and order the hierarchical labels (Section, Class, Subclass, etc.) that lead to a specific code.

Memory-Moderate Questions - Path Reconstruction

Question: Give the name chain for F02B 19/00 from Section → Class → Subclass → Main group

Options:

- A) Mechanical Engineering → Pumps → Piston Engines → Precombustion Chambers
- B) Mechanical Engineering → Combustion Engines → Piston Engines → Engines with precombustion chambers
- C) Lighting → Engines → Combustion Engines → Precombustion Chambers
- D) Mechanical Engineering → Combustion Engines → Gas Turbines → Precombustion Chambers

Answer: B

Description: Engines with precombustion chambers

A.3.3 MEMORY-HEAVY (5 OR MORE HIERARCHICAL RECALLS)

Cousin Relationship: This is a Memory-Heavy task that tests the model's ability to understand lateral relationships within the hierarchy. It requires traversing up to a common grandparent and then back down to identify a "first cousin" that shares the same ancestor.

Memory-Heavy Questions - Cousin Relationship

Question: First cousin of H04N 9/806 (same grandparent level) is:

Options:

- A) H04N 9/808
- B) H04N 9/815
- C) H04N 9/82
- D) H04N 9/804

Answer: B

Reasoning Paths:

- H04N 9/806 → H04N 9/804 (parent) → H04N 9/80 (grandparent) → H04N 9/808 (uncle/aunt)
- H04N 9/806 → H04N 9/804 (parent) → H04N 9/80 (grandparent) → H04N 9/81 (uncle/aunt) → H04N 9/815 (cousin)
- H04N 9/806 → H04N 9/804 (parent) → H04N 9/80 (grandparent) → H04N 9/82 (uncle/aunt)
- H04N 9/806 → H04N 9/804 (parent) → H04N 9/80 (grandparent) → H04N 9/804 (uncle/aunt)

Deepest Descent: This Memory-Heavy question asks the model to identify the most specific descendant of a given patent code from a list of options. It tests the model's ability to perform deep, multi-step traversal down the hierarchy to determine which option has the longest, most specific path.

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

Memory-Heavy Questions - Deepest Descent

Question: Most specific descendant of A01B 3/00 among these options:

Options:

- A) A01B 3/04
- B) A01B 3/426
- C) A01B 3/08
- D) A01B 3/26

Answer: B

Reasoning Paths:

- A01B 3/00 → A01B 3/04
- A01B 3/00 → A01B 3/36 → A01B 3/40 → A01B 3/42 → A01B 3/426
- A01B 3/00 → A01B 3/04 → A01B 3/06 → A01B 3/08
- A01B 3/00 → A01B 3/24 → A01B 3/26

Orphan Detection: As a Memory-Heavy task, this question asks the model to identify which of the given code pairs does not represent a valid parent-child relationship. This tests the model's deep knowledge of the hierarchical structure and its ability to spot inconsistencies.

Memory-Heavy Questions - Orphan Detection

Question: Which does not represent a valid parent-child relationship?

Options:

- A) D01F 6/26 → D01F 6/28
- B) D01G 19/14 → D01G 19/16
- C) D01B 5/02 → D01B 5/04
- D) D01D 1/06 → D01D 1/09

Answer: A

Reasoning:

- D01B 1/14 → D01B 1/18
- D01F 2/24 → D01F 2/28 → D01F 2/30 (not D01F 2/06)
- D01G 15/76 → D01G 15/78 (not D01G 15/74)
- D01D 5/04 is parallel to D01D 5/08, not a child

Common Ancestor: This task can range from Memory-Light to Memory-Heavy, depending on the codes provided. It requires the model to navigate the hierarchical paths of two different codes to find their nearest shared ancestor, testing both traversal and comparison skills. Below is an example showing the highest retrieval complexity:

Memory-Heavy Questions - Common Ancestor

Question: Nearest common ancestor of A01B 3/421 and A01B 15/06 is:

Options:

- A) A01B 3/00
- B) A01B 15/00
- C) A01B
- D) A01

Answer: C

Hierarchical Paths:

- A01B 3/421 → A01B 3/42 → A01B 3/40 → A01B 3/36 → A01B 3/00 → A01B
- A01B 15/06 → A01B 15/04 → A01B 15/02 → A01B 15/00 → A01B

B ADDITIONAL RESULTS

B.1 STRUCTURED PROMPTING AND MODEL PERFORMANCE

Figure 3 presents dumbbell plots that illustrate how structured prompting can narrow the performance gap between base and enhanced language models on hierarchical classification tasks. The plots compare accuracy across different model families and enhancement categories, such as instruction-tuned and reasoning-enhanced models. The key finding is that using structured prompts effectively reduces the performance advantage of specialized models, suggesting prompt engineering can be a powerful alternative to other methods like reinforcement learning.

C USE OF LLMs

We made use of LLMs to perform tasks including the polishing of our writing and the LaTeX formatting of the manuscript. It is important to note that no new ideas were introduced by these models. The sole exception is where the focus of our research, prompt optimization, made us use the extraction and analysis of the LLMs' own responses.

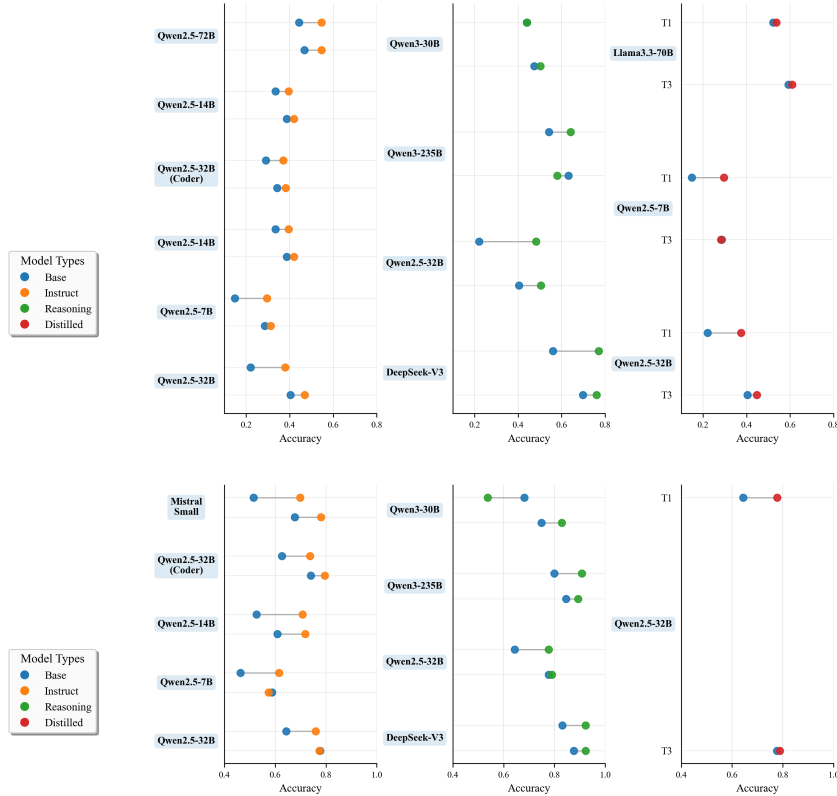


Figure 3: **Dumbbell plots revealing performance gap compression through structured prompting.** Accuracy comparison between base and enhanced models on (a) MedConceptsQA and (b) IPC codess across Templates 1 and 3. Each dumbbell connects base model performance (left endpoint) to enhanced model performance (right endpoint), with line length representing the performance gap. Model pairs span three enhancement categories: *Instruction-tuned* (Qwen2.5-7B/14B/32B/72B→Instruct, Qwen2.5-32B→Coder, Mistral-Small→Instruct), *Reasoning-enhanced* (Deepseek-V3→R1, Qwen2.5-32B→QwQ-32B, Qwen3-30B-A3B-Instruct→Thinking-2507, Qwen3-235B-A22B-Instruct→Thinking-2507), and *Distilled* (Qwen2.5-7/32B→R1-Distill-Qwen-7/32B, Llama3.3-70B-Instruct→R1-Distill-Llama-70B). The systematic compression of dumbbells from Template 1 to Template 3 demonstrates how structured prompting narrows or eliminates performance advantages of specialized models. Notably, several base models with Template 3 achieve parity or exceed their enhanced counterparts’ Template 1 performance, validating prompt engineering as an alternative to RL.