

# Experiments with different LMMs

## 1. MythoMax 13B (nitro)

Code: [gryphe/mythomax-l2-13b:nitro](#)

Description: One of the highest performing and most popular fine-tunes of Llama 2 13B, with rich descriptions and **roleplay**.

## GPT prompt output:

Let's test

**Short Result: 0 passed, 9 failed**

---

### Long Result:

- TESTCASE-1:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

- TESTCASE-2:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

- TESTCASE-3:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

- TESTCASE-4:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

- TESTCASE-5:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

- TESTCASE-6:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

- TESTCASE-7:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Expected RESPONSE: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

- TESTCASE-8:

1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

## Claude prompt output:

Short Result: 0 passed, 9 failed

Long Result:

- TESTCASE-1:  
Based on the specifications provided, the correct Series option that represents cabinet materials and thickness is:  
1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core  
This option has a square flat edge high-pressure laminate front, high-pressure laminate ends, industrial particleboard core that is CARB compliant, and a flat (.020") PVC cabinet edgebanding. The edge style is also a square flat edge.  
Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core
- TESTCASE-2:  
Based on the available Series options and the specifications provided, the correct Series that represents the cabinet materials and thickness is:  
1. 1600-C - Wood Veneer Frts 3mm Wood Edge & HPL Ends IPB CARB Compliant Core  
This Series option includes wood veneer fronts with a 3mm wood edge and high-pressure laminate (HPL) ends on an industrial particleboard (IPB) core that is compliant with the Carbon Monoxide (CO) emissions standards set by the California Air Resources Board (CARB).  
Expected RESPONSE: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core
- TESTCASE-3:  
Based on the available Series options, the correct Series that represents cabinet materials and thickness is:  
1150-C - Sq Flat Edge HPL Frts & TF Ends IPB CARB Compliant Core  
This Series option features 3mm High Pressure Laminate (HPL) fronts, Thermally Fused Laminate (TF) ends, Industrial Particleboard (IPB) core, and Flat (.020") PVC Cabinet Edgebanding. The core is CARB Compliant.  
Expected RESPONSE: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core
- TESTCASE-4:  
Based on the available Series options, the correct Series that represents the cabinet materials and thickness is:  
1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core  
This Series option has a square flat edge high-pressure laminate front and high-pressure laminate ends with an industrial particleboard core that is CARB compliant. The cabinet thickness is not specified in the given options.  
Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core
- TESTCASE-5:  
Based on the specification, the correct Series option that represents cabinet materials and thickness is:  
{  
"series": "1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core"  
}  
This series features a square flat edge high-pressure laminate front and high-pressure laminate end, with an industrial particleboard core that is CARB compliant. The cabinet edges are also square and flat, and the thickness of the fronts is not specified in the given options.

Result:

Tried to answer in few sentence, without Chain of Thoughts. MythoMax has very bad results.

2. Reflection 70B (free)

Code: [mattshumer/reflection-70b:free](https://huggingface.co/mattshumer/reflection-70b:free)

Description: Reflection Llama-3.1 70B is trained with a new technique called Reflection-Tuning that teaches a LLM to detect mistakes in its reasoning and correct course. The model was trained on synthetic data.

## GPT prompt output:

### Short Result: 1 passed, 8 failed

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1150-C - Sq Flat Edge HPL Frts & TF Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1250-C

Result: **failed**

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1600-C - Wood Veneer Frts 3mm Wood Edge & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

## TESTCASE - 3 OUTPUT:

<thinking>

To solve this problem, I'll follow these steps:

1. Identify key information in the specification related to cabinet materials and thickness.
2. Compare this information to the criteria outlined in the instruction.
3. Match the identified information with the available Series options.
4. Eliminate options that do not meet the criteria.
5. Select the most appropriate Series option based on the analysis.

Let's go through this process:

1. Key information from the specification:
  - o Cabinet ends: 3/4" thick
  - o Cabinet backs: 1/4" prefinished Medium Density Fiberboard (MDF)
  - o Adjustable shelves: 3/4" thick
  - o Drawer members: 3/4" thick
  - o Drawer bottoms: 1/4" MDF
2. Comparing to the instruction criteria:

The instruction asks for a Series option that represents cabinet materials and thickness. We need to focus on the thickness of the cabinet ends and the material of the cabinet backs.
3. Matching with available Series options:

Looking at the options, we can see that most of them have:

  - o Ends: High Pressure Laminate (HPL) or Thermally Fused Laminate (TF)
  - o Core: Industrial Particleboard (IPB) or No Added Urea Formaldehyde Particleboard (NAUF PB)
  - o Edgebanding: Flat (.020") PVC Cabinet Edgebanding
4. None of the options explicitly mention 3/4" thickness for ends or 1/4" MDF for backs. However, we need to find the closest match.
5. Eliminating options:

We can eliminate options that don't match the general criteria:

  - o Series 1600-C and 1600-N can be eliminated as they have wood veneer fronts, which is not mentioned in the specification.
  - o Series 1100-C, 1100-N, 1150-C, and 1150-N can be eliminated as they have square flat edge HPL fronts, which doesn't match the 3mm radius edge mentioned in the specification.
6. Selecting the most appropriate option:

After elimination, we're left with Series 1200-C, 1200-N, 1250-C, and 1250-N. These all have 3mm HPL fronts and 3mm radius edges, which seems to be the closest match to the specification.

Among these, Series 1200-C and 1200-N have HPL ends, while Series 1250-C and 1250-N have TF ends. The specification doesn't explicitly mention the type of ends, so we can't make a distinction based on this.

The main difference between these options is the core material:

  - o Series 1200-C and 1250-C have Industrial Particleboard (IPB) CARB Compliant Core
  - o Series 1200-N and 1250-N have No Added Urea Formaldehyde Particleboard (NAUF PB) Core
7. The specification doesn't mention CARB compliance or NAUF, so we can't make a decision based on this.

<reflection>

Upon reflection, I realize that I may have been too quick to eliminate the 1600 series options. While they do have wood veneer fronts, the specification doesn't explicitly rule out wood veneer. Additionally, these options have 3mm wood edges, which could be considered similar to the 3mm radius edge mentioned in the specification. Let's reconsider these options.

</reflection>

Reconsidering the 1600 series options:

- Series 1600-C and 1600-N have wood veneer fronts with 3mm wood edges, which could be seen as similar to the 3mm radius edge mentioned in the specification.
- They also have HPL ends, which is consistent with the specification's mention of 3/4" thick ends (though the exact thickness isn't specified in the options).
- The core material is IPB CARB Compliant Core for 1600-C and NAUF PB Core for 1600-N, which is similar to the other options.

Given this new consideration, Series 1600-C and 1600-N seem to be strong contenders. They match the 3mm edge requirement and have HPL ends, which is closer to the 3/4" thickness mentioned in the specification than the other options.

<reflection>

After this reflection, I believe Series 1600-C or 1600-N would be the best choice. They match the 3mm edge requirement and have HPL ends, which is closer to the 3/4" thickness mentioned in the specification than the other options. The wood veneer fronts could be seen as a minor deviation from the exact specification, but it's the closest match overall.

</reflection> </thinking> <output>

Based on the analysis of the specification and the available Series options, the most appropriate Series option that represents cabinet materials and thickness is either Series 1600-C or Series 1600-N.

These options stand out for several reasons:

1. Edge style: Both Series 1600-C and 1600-N have a 3mm radius wood edge, which closely matches the 3mm radius edge mentioned in the specification.
2. End material: They have HPL (High Pressure Laminate) ends, which is consistent with the specification's mention of 3/4" thick ends. While the exact thickness isn't specified in the options, HPL ends are a closer match to the 3/4" thickness requirement than the other options.
3. Core material: Both options have a particleboard core, with Series 1600-C having an Industrial Particleboard (IPB) CARB Compliant Core and Series 1600-N having a No Added Urea Formaldehyde Particleboard (NAUF PB) Core. This is similar to the core materials mentioned in the other options.
4. Overall construction: The wood veneer fronts with 3mm wood edges in these options could be seen as a minor deviation from the exact specification, but they provide a strong and durable construction that aligns with the overall quality implied in the specification.

While the other options (Series 1200-C, 1200-N, 1250-C, and 1250-N) have 3mm HPL fronts and 3mm radius edges, they don't match the end thickness requirement as closely as the 1600 series options. The 1100 and 1150 series options can be eliminated due to their square flat edge HPL fronts, which don't match the 3mm radius edge mentioned in the specification.

Therefore, between Series 1600-C and 1600-N, the choice would depend on whether CARB compliance (1600-C) or NAUF (1600-N) is preferred or required for the project. Both options represent the cabinet materials and thickness most closely based on the provided specification and available Series options.

</output>

Expected RESPONSE: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

## Result:

Reflection LLMs aim to explain each step of the instructions comprehensively but face significant challenges in executing nearly every step. This difficulty might stem from the specific synthetic data on which the model was trained. Despite its structured approach, the model often fails to accurately address key details outlined in the instructions. Notably, the model exhibits high determinism, consistently providing the same result across multiple prompts with zero temperature.

## Claude prompt output:

### Prompt test

#### Test your prompt

Choose LLM model:  
mattshumer/reflection-70b.free

Choose prompt:  
Prompt 2

Choose testcases:  
(1) '1200-N - 3m... x (2) '1100-C - Sq... x (3) '2800-C - 3m... x (4) '1200-N - 3m... x (5) '1200-C - 3m... x (6) '1200-N - 3m... x (7) '2200-C - 3m... x (8) '1250-C - 3m... x (9) '2100-C - 5q... x

Let's test

**Short Result: 0 passed, 9 failed**

---

## TESTCASE - 1 OUTPUT:

<thinking>

To solve this problem, I'll follow these steps:

1. Analyze the specification to identify key information about cabinet materials and thickness.
2. Compare this information with the criteria in the instruction.
3. Match the identified information with the available Series options.
4. Eliminate options that don't meet the criteria.
5. Select the most appropriate Series option.

Let's go through this process:

1. Analyzing the specification:
  - o Cabinet construction: Particleboard, plywood, hardwood lumber, or high-density fiber board.
  - o Thickness: Minimum 1/2-inch for sides, back, and sub-front of drawers.
  - o Bottoms: Minimum 1/4-inch tempered hardboard.
  - o Drawer bottoms: Minimum 1/2-inch thick particleboard.
2. Comparing with the instruction criteria:
  - o The instruction asks for a Series option that represents cabinet materials and thickness.
  - o We need to find an option that matches the materials and thicknesses specified.
3. Matching with available Series options:
  - o The options mention various materials like HPL (High Pressure Laminate), TF (Thermally Fused Laminate), and wood veneer.
  - o Thicknesses mentioned include 3mm (which is approximately 1/8 inch) and no specific thickness for wood veneer.
4. Eliminating options:
  - o Options with 3mm thickness (1100-C, 1100-N, 1150-C, 1150-N, 1200-C, 1200-N, 1250-C, 1250-N) can be eliminated as they don't meet the minimum 1/2-inch thickness requirement.
  - o Options with HPL or TF for all parts (1100-C, 1100-N, 1150-C, 1150-N, 1200-C, 1200-N, 1250-C, 1250-N) can be eliminated as they don't match the variety of materials specified in the instruction.
5. Selecting the most appropriate option:
  - o The only option that remains is 1600-C or 1600-N, as they are the only ones that mention wood veneer and have a 3mm wood edge, which could potentially meet the thickness requirement.

<reflection>

Upon reflection, I realize that I may have been too hasty in eliminating the 3mm options. The specification mentions a minimum 1/2-inch thickness for some parts, but not all. The 3mm thickness could potentially apply to certain components. However, the main issue is that none of the 3mm options mention the variety of materials specified in the instruction. The 1600-C and 1600-N options are still the most likely candidates as they mention wood veneer and have a 3mm wood edge, which could meet the thickness requirement for some parts.

</reflection>

After careful consideration, the most appropriate Series option appears to be either 1600-C or 1600-N. Both of these options mention wood veneer with a 3mm wood edge, which could potentially meet the thickness requirement for some parts of the cabinet. They also mention HPL for the ends, which aligns with the variety of materials specified in the instruction.

</thinking> <output>

Based on the analysis of the specification and the available Series options, the most appropriate Series option that represents cabinet materials and thickness is either 1600-C or 1600-N.

These options are the best match for several reasons:

1. Material variety: They mention wood veneer for the fronts, which aligns with the specification's mention of various materials including wood veneer.
2. Thickness: While the specification mentions a minimum 1/2-inch thickness for some parts, the 3mm wood edge mentioned in these options could potentially meet this requirement for certain components.
3. Core material: Both options mention Industrial Particleboard CARB Compliant Core (1600-C) or No Added Urea Formaldehyde Particleboard (1600-N), which aligns with the specification's mention of particleboard as a possible material.
4. Edge style: The 3mm Radius Wood Edge mentioned in these options could potentially meet the thickness requirement for some parts of the cabinet.
5. Overall construction: These options seem to offer a more comprehensive construction method that could accommodate the various materials and thicknesses specified in the instruction.

It's important to note that while these options are the best match, they don't perfectly align with all aspects of the specification. The specification mentions a minimum 1/2-inch thickness for some parts, which is not explicitly stated in these options. However, given the available choices, 1600-C or 1600-N appear to be the most suitable options for representing cabinet materials and thickness as specified in the instruction.</output>

**Expected RESPONSE: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core**

## **Result:**

Reflection LLM follows the prompt structure closely, striving to execute every step of the provided instructions and guidelines. However, it often struggles with selecting the right keywords and sentences from specification and has difficulty identifying the key elements to focus on when analyzing the cabinet specification.

## **3. Meta: Llama 3.1 8B Instruct (free)**

Code: [meta-llama/llama-3.1-8b-instruct:free](#)

### **Description:**

Meta's latest class of model (Llama 3.1) launched with a variety of sizes & flavors. This 8B instruct-tuned version is fast and efficient. It has demonstrated strong performance compared to leading closed-source models in human evaluations.



## GPT prompt output:

**Short Result: 0 passed, 9 failed**

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C

Result: **failed**

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1150-N - Sq Flat Edge HPL Frts & TF Ends NAUF PB Core

Result: **failed**

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-C

Result: **failed**

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C

Result: **failed**

- TESTCASE-6: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1600-C

Result: **failed**

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C

Result: **failed**

---

## Claude prompt output:

## Short Result: 0 passed, 9 failed

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1600-C - Wood Veneer Frts 3mm Wood Edge & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-6: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: Series 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: **failed**

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1250-C

Result: **failed**

.....

## TESTCASE - 8 OUTPUT:

To follow the instructions and pick the correct series option that represents the cabinets' materials and thickness, I need to identify the key characteristics mentioned in the specification.

From section "A. 3/4-inch panel construction" and "B. Cabinet Frame Construction", I understand that the cabinets have:

- 3/4-inch thickness for base cabinets, including ends, bottoms, and top and back frames
- Dowel pin construction for stability
- No separate attached bases
- Integrated design with back groove and precision CNC drill pattern for accurate location of fixed members, hardware, and shelf supports

From section "C. Back Procurement of Materials", I see that the specific details mentioned include:

- Backs shall be an integrated system of a 1/4-inch prefinished Medium Density Fiberboard (MDF) captured in side and horizontal grooves

- Unit back to be further integrated with attachment to 3/4-inch doweled-in mounting frames
- Fixed backs are mechanically fastened into grooves and sealed with hot melt adhesive

Considering the specification requirements, I will look for the series that matches these conditions.

The options that best match these specifications are:

- Series 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core (does not match the required 3/4-inch thickness)
- Series 1600-C - Wood Veneer Frts 3mm Wood Edge & HPL Ends IPB CARB Compliant Core (does not match the required 3/4-inch thickness)

After analyzing the options and understanding that series 1100 and 1150 are plain 3/4-inch Flat Edge (does not match the acceptance of dowel pin construction) and series 1200 and 1250 provide 3mm thickness (does not match the required 3/4-inch thickness), so the series must be at least 3/4 inch

However, upon closer inspection of the specification for backs, an important aspect is that the backs have a prefinished 1/4-inch MDF. However, there is also an important aspect of dowel pin construction that requires the integral design (series:\*. The 1600 series provides a 3mm Wood Veneer and integrated back design.

However, from the specified options none match fully.

series "1150-C" option shares the integral back feature and the dowel pin construction and 3/4 thickness.

ANSWER: series "1150-C"

Expected RESPONSE: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

## Result:

**The LLaMA 3.1 8B model tends to provide very brief and short responses and struggles with effectively utilizing the "chain of thought" technique, often making non-transparent decisions that can be difficult to follow. Also LLaMA 3.1 8B has very low determinism and very often writes different solutions for the same test case.**

## 4. Meta: Llama 3.1 405B Instruct

**Description:** The highly anticipated 400B class of Llama3 is here! Clocking in at 128k context with impressive eval scores, the Meta AI team continues to push the frontier of open-source LLMs.

Meta's latest class of model (Llama 3.1) launched with a variety of sizes & flavors. This 405B instruct-tuned version is optimized for high quality dialogue use cases.

Code: [meta-llama/llama-3.1-405b-instruct](#)

## GPT prompt output:

meta-llama/llama-3.1-405b-instruct

Choose prompt:  
Prompt 1

Choose testcases:  
[1], 1200-N - 3m... x [2], 1100-C - 5q... x [3], 2800-C - 3m... x [4], 1250-C - 3m... x [9], 2100-C - 5q... x [7], 2200-C - 3m... x

Let's test

**Short Result: 3 passed, 3 failed**

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core  
LLM RESPONSE: RESULT: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core  
Result: success
- TESTCASE-2: 1100-C - 5q Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1250-N - 3mm HPL Frts & TF Ends NAUF PB Core  
Result: failed
- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core  
Result: failed
- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core  
Result: success
- TESTCASE-9: 2100-C - 5q Flat Edge TF Frts & Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core  
Result: failed
- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core  
Result: success

.....

**Long Result:**

TESTCASE-1:

## Claude prompt output:

Choose LLM model:  
meta-llama/llama-3.1-405b-instruct

Choose prompt:  
Prompt 2

Choose testcases:  
 (1', '1200-N - 3m... x (2', '1100-C - Sq... x (3', '2800-C - 3m... x (8', '1250-C - 3m... x (9', '2100-C - Sq... x (7', '2200-C - 3m... x

Let's test

**Short Result: 4 passed, 2 failed**

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core  
LLM RESPONSE: RESULT: 1250-N - 3mm HPL Frts & TF Ends NAUF PB Core  
Result: failed
- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core  
Result: failed
- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core  
Result: success
- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core  
Result: success
- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core  
Result: success
- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core  
LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core  
Result: success

.....  
Long Result:

**Result:**

The *Meta: LLaMA 3.1 405B Instruct* model exhibited a well-structured approach to problem-solving and achieved a decent percentage of correctly executed test cases. However, its success rate was notably lower compared to models like GPT and Claude, which outperformed LLaMA in terms of accuracy and handling complex, multi-step tasks. Despite this, LLaMA still demonstrated promise in maintaining logical consistency and structure in its responses.

**5. Anthropic: Claude 3.5 Sonnet (self-moderated)**

Code: [anthropic/claude-3.5-sonnet:beta](#)

## Claude prompt output:

### Short Result: 7 passed, 2 failed

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: **success**

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: **success**

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-6: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **failed**

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

## TESTCASE - 8 OUTPUT of Claude-3-opus:

<thinking>

### 1. Core Material Selection:

The specification mentions particleboard core materials with a minimum density of 45-48 lb, tested under ANSI A 208.1 and/or ASTM D-1037-91A standards. It does not explicitly prohibit adhesives containing urea formaldehyde. Therefore, I will select Industrial Particleboard CARB Compliant as the core material.

### 2. Fronts Material Selection:

The specification states that door and drawer fronts must be matching high pressure laminate on both sides. Therefore, I will select 3mm High Pressure Laminate as the fronts material.

### 3. Fronts Thickness Determination:

The specification mentions that door and drawer fronts shall be 3/4" thick. This aligns with the standard thickness for 3mm High Pressure Laminate fronts.

### 4. Ends Material Selection:

The specification does not explicitly mention the material for cabinet ends. However, it states that all exposed/semi-exposed surfaces shall be color matching thermofused or high pressure laminate. Since the fronts are specified as HPL, I will select High Pressure Laminate as the ends material to maintain consistency.

### 5. Fronts Edging Selection:

The specification clearly states that door and drawer front edges shall have 3mm radius extrusion banding. Therefore, I will select 3mm Radius Edge as the fronts edging style.

### 6. Cabinet Edgebanding Selection:

The specification mentions that cabinet sides, top, bottom, adjustable shelves, and other interior components shall be edged with (.020") flat edge extrusion. This matches the default option of Flat (.020") PVC Cabinet Edgebanding.

Based on the provided specification and the selection criteria, the most appropriate Series option is:

5120-C - 3mm HPL Frts & Ends IPB CARB Compliant Core

</thinking>

<result>

RESULT: 5120-C - 3mm HPL Frts & Ends IPB CARB Compliant Core

</result>

**Pricing for claude-3.5-sonnet (self-moderated the same)**

Timestamp	Model	App	Tokens ⌵	Cost ⌵	Speed ⌵	Provider ⌵
Sep 17, 12:48:29 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	8611 → 488	\$ 0.0332	52.3 tps	Anthropic >
Sep 17, 12:48:29 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	9889 → 518	\$ 0.0374	57.0 tps	Anthropic >
Sep 17, 12:48:27 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	17274 → 409	\$ 0.058	58.5 tps	Anthropic >
Sep 17, 12:48:27 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	7945 → 371	\$ 0.0294	48.7 tps	Anthropic >
Sep 17, 12:48:27 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	9859 → 439	\$ 0.0362	54.9 tps	Anthropic >
Sep 17, 12:48:27 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	7584 → 488	\$ 0.0301	62.8 tps	Anthropic >
Sep 17, 12:48:27 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	7423 → 421	\$ 0.0286	55.7 tps	Anthropic >
Sep 17, 12:48:26 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	9358 → 443	\$ 0.0347	57.2 tps	Anthropic >
Sep 17, 12:48:26 PM	<a href="#">Claude 3.5 Sonnet</a> ⌵	Unknown	7890 → 406	\$ 0.0298	56.6 tps	Anthropic >

**Cost for 9 test cases: 0.32 \$**

**Average cost for 1 testcase: 0.036 \$**

### For example pricing for gpt-4o-2024-05-13

Timestamp	Model	App	Tokens ⌵	Cost ⌵	Speed ⌵	Provider ⌵
Sep 18, 10:51:15 AM	<a href="#">GPT-4o</a> ⌵	Unknown	6926 → 773	\$ 0.0462	69.7 tps	OpenAI >
Sep 18, 10:51:12 AM	<a href="#">GPT-4o</a> ⌵	Unknown	6167 → 602	\$ 0.0399	64.2 tps	OpenAI >
Sep 18, 10:51:11 AM	<a href="#">GPT-4o</a> ⌵	Unknown	6356 → 506	\$ 0.0394	64.7 tps	OpenAI >
Sep 18, 10:51:11 AM	<a href="#">GPT-4o</a> ⌵	Unknown	7906 → 458	\$ 0.0464	59.6 tps	OpenAI >
Sep 18, 10:51:10 AM	<a href="#">GPT-4o</a> ⌵	Unknown	5797 → 360	\$ 0.0344	58.3 tps	OpenAI >
Sep 18, 10:51:09 AM	<a href="#">GPT-4o</a> ⌵	Unknown	7878 → 341	\$ 0.0445	57.2 tps	OpenAI >
Sep 18, 10:51:09 AM	<a href="#">GPT-4o</a> ⌵	Unknown	7566 → 350	\$ 0.0431	61.3 tps	OpenAI >
Sep 18, 10:51:08 AM	<a href="#">GPT-4o</a> ⌵	Unknown	14071 → 349	\$ 0.0756	71.2 tps	OpenAI >
Sep 18, 10:51:08 AM	<a href="#">GPT-4o</a> ⌵	Unknown	6048 → 307	\$ 0.0348	64.1 tps	OpenAI >

**Cost for 9 test cases: 0.41 \$**

**Average cost for 1 testcase: 0.046 \$**



**Result:**

Models such as *anthropic/claude-3-opus* and *anthropic/claude-3-sonnet(self-moderated)* also delivered highly accurate results, demonstrating a high level of determinism and a structured approach to executing multi-step complex tasks. Their responses aligned consistently with the expected output format, particularly when generating structured responses typical of *sonnet* output. This consistency in following intricate steps showcases their capability to handle detailed instructions with precision and reliability.

**6. Qwen 2 72B Instruct**

Code: [qwen/qwen-2-72b-instruct](#)

## GPT prompt output:

**Short Result: 8 passed, 1 failed**

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series 1200-N

Result: **success**

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: series

Result: **success**

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: series "2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core"

Result: **success**

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series "1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core"

Result: **success**

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-6: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: **success**

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

Result: **success**

- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: series

Result: **failed**

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: - RESULT: series "1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core"

Result: **failed**

## Claude prompt output:

Let's test

## Short Result: 4 passed, 5 failed

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1250-N - 3mm HPL Frts & TF Ends NAUF PB Core

Result: failed

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1100-N - Sq Flat Edge HPL Frts & HPL Ends NAUF PB Core

Result: failed

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: The selected Series option is 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core.

Result: failed

- TESTCASE-3: 2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: failed

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: success

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: success

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

Result: success

- TESTCASE-8: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1250-C - 3mm HPL Frts & TF Ends IPB CARB Compliant Core

Result: success

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

### Short Result: 5 passed, 4 failed

- TESTCASE-1: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series 1200-N

Result: failed

- TESTCASE-2: 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: series = 1100-C - Sq Flat Edge HPL Frts & HPL Ends IPB CARB Compliant Core

Result: success

- TESTCASE-3: 2800-C - 3mm Frts & TF CoLor Throughout IPB CARB Compliant Core

LLM RESPONSE: RESULT: series "2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core"

Result: success

- TESTCASE-4: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series = 1200-N

Result: failed

- TESTCASE-5: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-C - 3mm HPL Frts & HPL Ends IPB CARB Compliant Core

Result: success

- TESTCASE-6: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

LLM RESPONSE: RESULT: series "1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core"

Result: success

- TESTCASE-7: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 2200-C - 3mm TF Frts & Ends IPB CARB Compliant Core

Result: success

- TESTCASE-9: 2100-C - Sq Flat Edge TF Frts & Ends IPB CARB Compliant Core

LLM RESPONSE: RESULT: 1200-N - 3mm HPL Frts & HPL Ends NAUF PB Core

Result: failed

---

Long Result:

## TESTCASE - 3 GPT PROMPT:

### Justification:

1. Core Material Selection: The specification mentions particleboard core materials with a minimum density of 45-48 lb. Since there's no mention of adhesives containing urea formaldehyde, we default to "Industrial Particleboard CARB Compliant."
2. Fronts Material Selection: The specification requires door and drawer fronts to be finished entirely in matching high pressure laminate (HPL) materials on both the fronts and backs. The rest of the exposed/semi-exposed cabinet components can be either color matching thermofused (TFL) or matching HPL. Since the fronts must be HPL, we select "3mm High Pressure Laminate."
3. Fronts Thickness Determination: The specification indicates that door and drawer fronts shall be 3/4" thick. This corresponds to the "3mm" thickness option for fronts.
4. Ends Material Selection: The specification allows for either color matching TFL or matching HPL for the rest of the exposed/semi-exposed cabinet components. Since we have already selected HPL for the fronts, we default to HPL for the ends.
5. Fronts Edging Selection: The specification requires 3mm radius extrusion banding for door and drawer fronts. This corresponds to the "3mm Radius Edge" style.
6. Cabinet Edgebanding Selection: The specification requires (.020") flat edge extrusion for cabinet sides, top, bottom, adjustable shelves, and other interior components. This corresponds to the "Flat (.020") PVC Cabinet Edgebanding" option.

The ""2800-C - 3mm Frts & TF Color Throughout IPB CARB Compliant Core"" series matches all the requirements specified in the input text. It features a 3mm High Pressure Laminate for fronts, matching High Pressure Laminate for ends, an Industrial Particleboard CARB Compliant core, and a 3mm Radius Edge style for the fronts, with Flat (.020"" ) PVC Cabinet Edgebanding for the cabinet edges.

### Pricing for [qwen/qwen-2-72b-instruct](#)

Sep 17, 12:33:13 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	6058 → 460	\$ 0.0023	32.9 tps	DeepInfra >
Sep 17, 12:33:02 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	6441 → 435	\$ 0.00243	32.9 tps	DeepInfra >
Sep 17, 12:33:01 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	8297 → 399	\$ 0.00306	32.9 tps	DeepInfra >
Sep 17, 12:32:59 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	7908 → 353	\$ 0.00291	33.2 tps	DeepInfra >
Sep 17, 12:32:59 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	8295 → 28	\$ 0.00291	31.4 tps	DeepInfra >
Sep 17, 12:32:59 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	6227 → 352	\$ 0.00232	33.1 tps	DeepInfra >
Sep 17, 12:32:58 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	14724 → 417	\$ 0.0123	54.3 tps	Together >
Sep 17, 12:32:49 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	7199 → 27	\$ 0.00253	29.3 tps	DeepInfra >
Sep 17, 12:32:49 PM	<a href="#">Qwen 2 72B Instruct</a> ▾	Unknown	6587 → 26	\$ 0.00232	29.4 tps	DeepInfra >

**Cost for 9 test cases: 0.04 \$**

**Average cost for 1 testcase: 0.0045 \$**

#### RESULT:

#### Performance of Qwen 2 72B Instruct in Task Execution

The *Qwen 2 72B Instruct* model demonstrated impressive results on prompts designed for GPT, successfully completing 8 out of 9 tasks. However, this success was achieved when the outputs were manually reviewed rather than programmatically analyzed, due to the model's poorly formatted responses. Despite the formatting issues, the answers were often correct, which is remarkable. It also exhibited a moderate level of determinism, and the high task completion rate suggests it is well-suited for executing such complex tasks.

With a cost of **\$0.35/M** input tokens and **\$0.4/M** output tokens (0.04 \$ for 9 test cases), **Qwen-2-72B** presents itself as a viable option for further testing, when:

- **gpt-4o-2024-08-06** \$2.5/M input and \$10/M output (0.21 \$ for 9 test cases)
- **gpt-4o-2024-05-13** \$5/M input and \$10/M output (0.40 \$ for 9 test cases)

- **claude-3.5-sonnet** \$3/M input and \$15/M output (0.32 \$ for 9 test cases)