***Classification of Events***

**True Positives (TP) - Correct Predictions by LLM.** These events were **correctly predicted by the LLM** and also happened in the real Battle of Normandy.

| **Event** | **Definition** |
| --- | --- |
| **Defenders had strong fortifications** | The German Atlantic Wall was fortified with bunkers, artillery, machine-gun nests, and beach obstacles. |
| **Defenders had elite troops and experienced leadership** | The German Army included well-trained forces (e.g., Waffen-SS, Wehrmacht) under commanders like Rommel and Rundstedt. |
| **Defenders suffered from fuel shortages** | Germany had major fuel shortages, limiting tank mobility and logistics. |
| **Defenders' supply lines were disrupted** | Allied bombings and sabotage significantly hindered German transportation. |
| **Defenders misinterpreted the invasion location** | Hitler and German High Command wrongly believed the attack would happen at Pas-de-Calais. |
| **Attackers had superior naval and air power** | The Allies controlled air and sea, limiting German counterattacks. |
| **Attackers had strong logistics and supply lines** | The Allies had robust supply chains, including artificial harbors (*Mulberry Harbors*) and fuel pipelines (*PLUTO*). |
| **Attackers faced difficult landing terrain** | The Normandy beaches were fortified with obstacles, making landings difficult. |
| **Weather played a role in the battle** | D-Day was postponed from June 5 to June 6 due to weather. |
| **Attackers suffered heavy casualties** | Omaha Beach saw particularly high Allied casualties due to heavy German resistance. |
| **Deception was used to mislead defenders** | *Operation Fortitude* created a fake army at Pas-de-Calais, tricking German intelligence. |
| **Defenders' delayed response led to their failure** | Hitler’s command delays and misinterpretation of the invasion site slowed German reinforcements. |
| **Air and naval dominance was key to victory** | The Allies' constant bombardments and air cover prevented German troop movements. |
| **Defenders' morale weakened over time** | Many German units, particularly conscripts, had low morale and surrendered. |
| **German industrial and supply chain collapse followed the battle** | Allied bombings severely impacted Germany’s ability to resupply its forces. |
| **International pressure influenced the battle** | The invasion had major political significance and was coordinated with the Soviet Union to relieve pressure on the Eastern Front. |
| **The battle turned the war in favor of the Allies** | D-Day marked the beginning of the end for Nazi Germany. |

Total True Positives (TP): **17**

**False Positives (FP) - Incorrect Predictions by LLM.** These are events that the LLM **predicted, but did not actually happen** in the real Battle of Normandy.

| **Event** | **Why It Did NOT Happen?** |
| --- | --- |
| **Defenders launched large-scale counterattacks** | The Germans were slow to respond; no massive counterattack occurred. |
| **Attackers used cyber warfare for deception** | Cyber warfare did not exist in WWII; deception was done using radio signals and fake military units. |
| **Unexpected diplomatic ceasefire mid-battle** | There was no diplomatic pause; fighting continued until objectives were secured. |
| **Attackers faced major internal command disagreements** | Allied leadership remained mostly united; there were no major command conflicts. |
| **Defenders received last-minute reinforcements from another front** | German reinforcements were blocked by destroyed bridges and delayed orders. |

Total False Positives (FP): **5**

**False Negatives (FN) - Missed Predictions by LLM.** These are **real events that happened in the Battle of Normandy but were NOT predicted** by the LLM.

| **Event** | **Definition** |
| --- | --- |
| **Allies used specialized engineering units to clear obstacles** | U.S. **Rangers** and British **Commandos** used **demolitions** to remove beach obstacles. |
| **Hedgerows slowed the Allies' inland advance** | Normandy’s **thick hedgerows (bocage)** slowed tanks and infantry movements post-landing. |
| **Hitler’s direct involvement delayed German counterattacks** | Hitler personally refused to release **Panzer divisions** early, believing the Normandy invasion was a distraction. |
| **Allies built artificial harbors (Mulberry Harbors)** | Allies created **temporary floating harbors** to offload supplies directly onto beaches. |
| **Allies used underwater pipelines (PLUTO) to transport fuel** | The **Pipeline Under the Ocean (PLUTO)** provided fuel from England to France. |
| **Rommel failed to convince Hitler about the real threat** | Rommel warned of an invasion at Normandy but was ignored. |
| **German command suffered from poor communication** | German **orders were delayed**, and poor coordination led to ineffective defenses. |
| **The Soviet Union’s offensive was timed with D-Day** | The **Soviets launched Operation Bagration** shortly after D-Day, pressuring Germany from the east. |

Total False Negatives (FN): **8**

***Final Results***

| **Metric** | **Value** |
| --- | --- |
| Total True Positives (TP) | 17 |
| Total False Positives (FP) | 5 |
| Total False Negatives (FN) | 8 |
| Precision | **77.27%** |
| Recall | **68.00%** |
| F1-Score | **72.10%** |

***Analysis and Insights***

**How well did the LLM predict the battle?**

* **Strengths:**
  + The **precision (77.27%)** is relatively high, meaning most of the LLM’s predictions were correct.
  + It accurately captured **key strategic and logistical factors** (fortifications, supply chain, deception, air/naval power).
  + It correctly predicted the **importance of morale and intelligence failures**.
* **Weaknesses:**
  + The **recall (68.00%)** indicates **some key details were missed**, such as:
    - The **use of engineering units** to remove obstacles.
    - The **importance of the bocage (hedgerows)** post-landing.
    - **Hitler’s delays and mismanagement**.
    - **The coordination with the Soviet Union**.
  + The **F1-score (72.10%)** suggests **solid but improvable performance**, especially in tactical details.

**What did the LLM get wrong?**

* It **invented events that did not happen**, such as:
  + A **large German counterattack**.
  + A **ceasefire or diplomatic pause** mid-battle.
  + **Cyber warfare**, which did not exist in WWII.
* It **missed some crucial aspects** like:
  + **Specialized Allied engineering efforts** (Rangers, PLUTO, Mulberry Harbors).
  + **Post-landing terrain challenges** (hedgerows).

***Final Verdict***

The LLM’s battle prediction was fairly accurate (~72.10% F1-score), but it missed some operational and strategic nuances of D-Day.  
It overestimated the defenders' ability to launch counterattacks and underestimated terrain-based challenges.  
Improving recall (identifying more historical details) would make future predictions more reliable.