***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 Baghdad (2003).**

| **Event** | **Definition** |
| --- | --- |
| **Invading forces had advanced tanks and armored vehicles** | The U.S.-led coalition used **M1 Abrams tanks, Bradley IFVs, and Stryker vehicles** to dominate Iraqi ground forces. |
| **Invading forces had complete air superiority** | The U.S. **Air Force dominated the airspace**, bombing strategic targets and neutralizing Iraq’s air defenses. |
| **Invading forces had strong logistics and supply chains** | U.S. forces had **pre-positioned fuel, ammunition, and food stockpiles**, allowing sustained operations. |
| **Invading forces led a coalition of allied nations** | The **U.S., U.K., Australia, and Poland** formed a coalition in the Iraq War. |
| **Defenders had outdated military technology** | Iraq relied on **Soviet-era T-72 tanks, old anti-aircraft systems, and limited night vision capabilities.** |
| **Defenders used urban guerrilla tactics** | Iraqi paramilitary forces and the Fedayeen Saddam used **hit-and-run ambushes** and **IEDs** in Baghdad. |
| **Defenders had supply chain disruptions** | U.S. airstrikes and ground offensives **destroyed Iraqi fuel depots and command centers** early in the battle. |
| **Defenders had morale and cohesion issues** | Many Iraqi **regular army units deserted** as the battle progressed, while **only elite Republican Guard divisions fought intensely.** |
| **Invading forces used "Shock and Awe" tactics** | The U.S. launched an **intense aerial bombardment** before ground forces attacked, intending to psychologically overwhelm Iraqi forces. |
| **Invading forces targeted infrastructure and key command centers** | The **U.S. bombed government buildings, power grids, and military headquarters** to disable Iraq’s operational capacity. |
| **Urban warfare created unexpected resistance pockets** | Some Iraqi fighters **held out longer than expected** in areas like Saddam International Airport and near the Tigris River. |
| **The battle resulted in the fall of Baghdad** | The city was captured by **April 9, 2003**, with the collapse of the Saddam Hussein government. |

Total True Positives (TP): **12**

**False Positives (FP) - Incorrect Predictions by LLM**

These are events that the LLM **predicted but did not actually happen** in the Battle of Baghdad.

| **Event** | **Why It Did NOT Happen?** |
| --- | --- |
| **Defenders had extensive fortifications** | Baghdad had **some defensive positions**, but not **fortified bunkers like in Normandy**. |
| **Defenders launched large-scale counterattacks** | The Iraqi military attempted **small-scale counterattacks**, but no coordinated large assault was ever launched. |
| **Unexpected diplomatic ceasefire mid-battle** | There was **no ceasefire**; the U.S. forces pushed continuously until victory. |
| **Cyber warfare played a significant role** | While some electronic warfare was used, **cyber attacks were not decisive in Baghdad.** |
| **Invaders faced major internal command disagreements** | The U.S. military command remained **largely unified in strategy and execution.** |

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

**False Negatives (FN) - Missed Predictions by LLM**

These are **real events that happened in the Battle of Baghdad but were NOT predicted** by the LLM.

| **Event** | **Definition** |
| --- | --- |
| **Iraqi forces disguised as civilians** | Iraqi fighters **dressed as civilians** and conducted surprise attacks, complicating U.S. rules of engagement. |
| **Saddam Hussein's disappearance created uncertainty** | The fall of Baghdad led to **Saddam going into hiding**, causing initial confusion over whether the war was over. |
| **U.S. forces used psychological warfare** | The coalition **dropped leaflets and used loudspeakers** to encourage Iraqi troops to surrender. |
| **Looting and lawlessness followed the battle** | After Baghdad fell, **mass looting** occurred, including the destruction of the **National Museum of Iraq**. |
| **Roadblocks and checkpoints slowed U.S. advance** | Iraqi forces placed **makeshift roadblocks and ambush points**, delaying the U.S. for short periods. |
| **The "Thunder Run" tactic broke through to Baghdad** | A **bold armored thrust (Thunder Run) led by U.S. tanks** pushed deep into Baghdad, securing the city faster than expected. |
| **U.S. faced early insurgent attacks post-battle** | Even after the fall of Baghdad, **insurgents immediately began ambushes, IED attacks, and sniper fire**. |
| **Media played a major role in shaping the battle’s perception** | Live coverage of the **toppling of Saddam’s statue** became an iconic moment of the invasion. |

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

***Final Results***

| **Metric** | **Value** |
| --- | --- |
| Total True Positives (TP) | **12** |
| Total False Positives (FP) | **5** |
| Total False Negatives (FN) | **8** |
| Precision | **70.59%** |
| Recall | **60.00%** |
| F1-Score | **64.62%** |

***Analysis and Insights***

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

* **Strengths:**
  1. Captured key strategic factors (air superiority, logistics, urban warfare).
  2. Correctly identified the major military asymmetry (U.S. technological advantage vs. outdated Iraqi forces).
  3. Recognized morale and supply chain issues on the defending side.
  4. Accurately described the use of "Shock and Awe" and infrastructure targeting.
* **Weaknesses:**
  1. **Recall was lower (60%)**, meaning **several important events were missed**, including:
     + The use of disguises by Iraqi fighters.
     + The importance of psychological warfare.
     + Looting and the breakdown of order after Baghdad fell.
  2. **Over-predicted** (False Positives) certain elements:
     + Baghdad was not fortified like Normandy.
     + There was no major Iraqi counterattack.
     + There was no cyber warfare component.

**What did the LLM get wrong?**

* Invented events that didn’t happen:
  + Massive fortifications (like D-Day) were not present in Baghdad.
  + Cyber warfare was not a decisive factor.
  + The Iraqi Army did not launch major counteroffensives.
* Missed some crucial aspects:
  + The "Thunder Run" tactic that secured the city.
  + The immediate insurgent attacks after Baghdad fell.
  + The role of media in shaping the battle’s perception.

***Final Verdict***

* The LLM’s battle prediction was decent (~64.62% F1-score), but it missed many operational and post-battle events.
* It overestimated the defenders' ability to fight conventionally and underestimated insurgent tactics.
* Improving recall (identifying more historical details) would make future predictions more reliable.