***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.

| **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. controlled 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, ensuring 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| Defenders had outdated military technology | Iraq relied on Soviet-era T-72 tanks, old anti-aircraft systems, and limited night vision capabilities.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 (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, while only elite Republican Guard divisions fought intensely. |
| Invading forces used "Shock and Awe" tactics | The U.S. launched an intense aerial bombardment before the ground invasion, aiming 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. |
| The battle resulted in the fall of Baghdad | The city was captured by April 9, 2003, leading to the collapse of Saddam Hussein's government. |

Total True Positives (TP): **11**

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

These are events that the LLM predicted, but did NOT actually happen in the real 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 occurred. |
| Cyber warfare played a major role | While some electronic warfare was used, cyber attacks were NOT decisive in Baghdad. |
| Unexpected diplomatic ceasefire mid-battle | There was no ceasefire; the U.S. pushed continuously until victory. |
| Invaders faced major internal command disagreements | The U.S. military command remained unified, and there were no significant leadership disputes. |

**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 U.S. movement for short periods. |
| The "Thunder Run" tactic broke through to Baghdad | A bold armored thrust (Thunder Run) led by U.S. tanks secured 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) | **11** |
| Total False Positives (FP) | **5** |
| Total False Negatives (FN) | **8** |
| Precision | **68.75%** |
| Recall | **57.89%** |
| F1-Score | **62.80%** |

***Critical Analysis of LLM’s Prediction***

1. Major Issue: The LLM Predicted the Wrong Winner

* The LLM incorrectly stated that the Defending Forces (Iraq) won when in reality, the Invading Forces (U.S.-led coalition) decisively won the Battle of Baghdad.
* This is a significant failure in outcome prediction.
* The U.S. captured Baghdad on April 9, 2003, and the Iraqi government collapsed—a fact the LLM did not correctly identify.

2. Strengths in LLM's Predictions

Correctly identified strategic factors:

* Air superiority was decisive.
* Iraq’s military was outdated and disorganized.
* Urban warfare caused resistance pockets.
* Coalition logistics and technology were superior.

Correctly predicted major tactics:

* "Shock and Awe" air campaign was central.
* Logistics and supply chains determined the battle.
* Defenders used guerrilla tactics and had morale issues.

3. Weaknesses in LLM’s Predictions

Failed to predict real-world insurgency impact:

* The U.S. faced immediate insurgent attacks post-victory, but the LLM missed this.

Did not account for looting and lawlessness:

* The destruction of Baghdad’s infrastructure after the battle was a major consequence not mentioned.

Overestimated defender resilience:

* Iraqi forces collapsed quickly, while the LLM predicted a prolonged resistance.

*Final Verdict*

* The LLM’s battle prediction was moderately accurate (~62.80% F1-score) but fundamentally flawed in outcome prediction.
* It overestimated the defenders' ability to fight conventionally and underestimated U.S. operational superiority.  
  Fixing recall errors (identifying more real-world details) and ensuring accurate outcome prediction would significantly improve future assessments.