### **Evaluation Summary – Harvey-Trained Model**

* **On Harvey (in-domain)**:  
   The model performs **exceptionally well**, with **high accuracy (96.7%)** and **strong F1-scores** across all categories, indicating it's well-tuned for the data it was trained on.
* **On Beryl (cross-domain)**:  
   Performance drops notably (**F1 ~ 52%**). While certain targets like tsaEligible remain strong, others such as roofDamage and replacementAssistanceEligible show **zero success.**
* **On Imelda (cross-domain)**:  
   Similar trends as Beryl — **moderate average accuracy (~55%)**, but some critical categories (tsaEligible, tsaCheckedIn) completely fail.

### **Evaluation Summary – Imelda-Trained Model**

* **On Imelda (in-domain)**:  
   The model performs **very well overall** with **96.8% average accuracy**. Most categories show strong precision and recall, except for tsaEligible and tsaCheckedIn, which have no positive examples (support = 0), so they don't contribute to F1. The model is clearly well-aligned with the Imelda data distribution.
* **On Harvey (cross-domain)**:  
   The average performance drops significantly (**F1 ~46%**). Categories like homeOwnersInsurance, floodInsurance, and floodDamage still show decent generalization. However, others such as destroyed, tsaEligible, and tsaCheckedIn fail completely.
* **On Beryl (cross-domain)**:  
   Results are **mixed**: some targets (homeOwnersInsurance, floodInsurance, floodDamage, rentalAssistanceEligible) perform excellently, even **better than on Harvey**, suggesting partial generalization. However, several key categories (destroyed, roofDamage, tsaEligible, tsaCheckedIn) again show **zero F1-score**, dragging the overall performance down (**F1 ~49%**).

### **Evaluation Summary – Beryl-Trained Model**

* **On Beryl (in-domain)**:  
   The model achieves **excellent performance**, with **~94% average accuracy** and **high F1-scores** across almost all targets. It handles categories like homeOwnersInsurance, tsaEligible, and floodDamage especially well.
* **On Harvey (cross-domain)**:  
   Performance is **moderate** with an average accuracy of ~64% and **F1 ~60%**. The model generalizes reasonably well to targets like tsaEligible, repairAssistanceEligible, and floodDamage. However, some key targets (roofDamage, replacementAssistanceEligible) fail completely.
* **On Imelda (cross-domain)**:  
   Results are **inconsistent**. While targets like homeOwnersInsurance and floodDamage still perform strongly, many others (roofDamage, tsaEligible, checkedIn, replacementAssistanceEligible) perform poorly or fail entirely — bringing the average F1 down to **~44%**.

### **Evaluation Summary – Combined Model (Harvey + Beryl + Imelda)**

* **On Beryl :**   
   The model delivers **excellent results**, with **~94% accuracy** and **F1 ≈ 84%**, closely matching performance of the Beryl-only trained model. Nearly all categories perform well, including those that usually struggle like replacementAssistanceEligible. However, roofDamage again suffers due to missing support.
* **On Harvey** :  
   Despite being part of the training set, the model's average performance is **moderate (~60% F1)**. It generalizes well for categories like tsaEligible, floodDamage, and rentalAssistanceEligible, but fails on others such as roofDamage and replacementAssistanceEligible. This suggests that **Harvey’s data patterns are harder to learn i**n the merged training set.
* **On Imelda** :  
   Similar to previous results, the model performs well on common categories (homeOwnersInsurance, floodDamage) but fails completely for tsaEligible, tsaCheckedIn, and roofDamage. These are consistent weak spots, even in multi-domain training. Average F1 is **~45%**, showing **limited generalization**.

### **Evaluation Summary – Combined Model (Harvey + Beryl)**

* **On Beryl (in-domain)**:  
   The model performs **extremely well**, matching the performance of the Beryl-only and 3-storm models, with **~93% accuracy** and **F1 ≈ 84%**. Nearly all categories show strong generalization. Some improvement is also seen in the replacementAssistanceEligible category, although roofDamage remains unsupported.
* **On Harvey (in-domain)**:  
   The model achieves **moderate success**, similar to previous multi-domain results, with **F1 ~61%**. High recall across most targets suggests the model is conservative and favors recall over precision. However, some targets like roofDamage and replacementAssistanceEligible perform poorly.
* **On Imelda (cross-domain)**:  
   The model struggles with Imelda. While a few categories such as homeOwnersInsurance, floodInsurance, and floodDamage transfer well, several important targets (tsaEligible, tsaCheckedIn, roofDamage) again show **zero performance**. This limits the average F1 to **~45%**.

### **Evaluation Summary – Combined Model (Harvey + Imelda)**

* **On Beryl (cross-domain)**:  
   The model **struggles significantly**, with an average **F1-score of ~41.6%**. While common categories like homeOwnersInsurance, floodInsurance, and floodDamage are handled decently, many targets — particularly destroyed, TSA-related ones, and replacementAssistanceEligible — fail completely.
* **On Harvey (in-domain)**:  
   Performance is **moderate** and comparable to previous Harvey evaluations with similar models, reaching **~61% F1 average**. The model handles several targets well (tsaEligible, floodDamage, repairAssistanceEligible), but again fails in areas like roofDamage and replacementAssistanceEligible.
* **On Imelda (in-domain)**:  
   Despite Imelda being part of the training set, the average **F1 is low (~45%)**. The model performs adequately on shared targets (homeOwnersInsurance, floodDamage), but **completely fails on TSA-related targets** and roofDamage, just like in cross-domain cases.

### **Evaluation Summary – Combined Model (Imelda + Beryl)**

* **On Beryl (in-domain)**:  
   The model shows **excellent performance**, with **~94% accuracy** and **F1 ≈ 83%**. Almost all categories, including complex ones like tsaEligible, floodDamage, and personalPropertyEligible, perform near perfectly. The only exception is roofDamage, which has no support in this test set and thus cannot be evaluated.
* **On Harvey (cross-domain)**:  
   The model achieves **moderate results**, with an average **F1 of ~60%**. It performs well for targets like tsaEligible, floodDamage, and rentalAssistanceEligible, but fails completely on roofDamage and replacementAssistanceEligible. This indicates that **Harvey-specific damage types and rare targets** are not well learned without Harvey in the training set.
* **On Imelda (in-domain)**:  
   Surprisingly, the performance on Imelda is **relatively weak** despite Imelda being part of training. While some common targets (homeOwnersInsurance, floodInsurance, floodDamage) are handled well, others such as tsaEligible, tsaCheckedIn, and roofDamage still receive **zero F1-scores**. This points to **class imbalance** during training that prevents generalization even in-domain.

## **Overall Evaluation Inference (Simplified)**

### **In-Domain Performance**

* **Beryl-trained model** performs almost perfectly on Beryl due to its smaller dataset.
* **Imelda-trained model** struggles on Imelda despite being trained on it, especially with TSA-related categories.
* **Harvey-trained model** performs very well on Harvey, showing it's well-aligned with its own data.

### **Cross-Domain Generalization**

#### **Beryl as Evaluation Set**

Most models do well on Beryl, even when Beryl isn't in training. This is likely because Beryl tweets are simpler, newer, and more consistent.

#### **Harvey as Evaluation Set**

Most models perform okay on Harvey — not too high, not too low. Harvey tweets have a balanced mix, making them moderately easy to generalize to.

#### **Imelda as Evaluation Set**

All models struggle on Imelda, even when Imelda is part of training. TSA-related and damage-type categories often fail completely. Imelda tweets are likely longer, noisier, or harder to interpret.

#### **Single-Storm Models :** Work great on their own data but fall short on other storms.

#### **Harvey + Beryl :** This combo gives the best overall balance across Beryl and Harvey.

#### **Harvey + Imelda :** Misses Beryl patterns and performs poorly on it.

#### **Imelda + Beryl :** Doesn’t handle Harvey well, likely due to its unique content.

#### **All Three Combined :** Boosts Beryl performance further but doesn’t fix Imelda issues.

### **Key Takeaways : Beryl** is easiest to predict, **Imelda** is the hardest . **TSA and RoofDamage** are tough to predict across all models.