Transformerアーキテクチャの代表的な種類

- プレフィックスデコーダ (Prefix Decoder)
  - ❖ デコーダオンリーモデルが入力テキストのより豊かな非因果的表現を構築できるように改良されたアーキテクチャ
  - ❖ 入力シーケンスが非因果的マスクを持つように（すなわち、過去のトークンに制限されない）、アテンションマスクが変更されている
  - ❖ prefixトークンに対する双方向の注意と、生成されたトークンに対する一方向の注意を可能にする
  - ❖ U-PaLMやGLM-130Bなどで採用されている

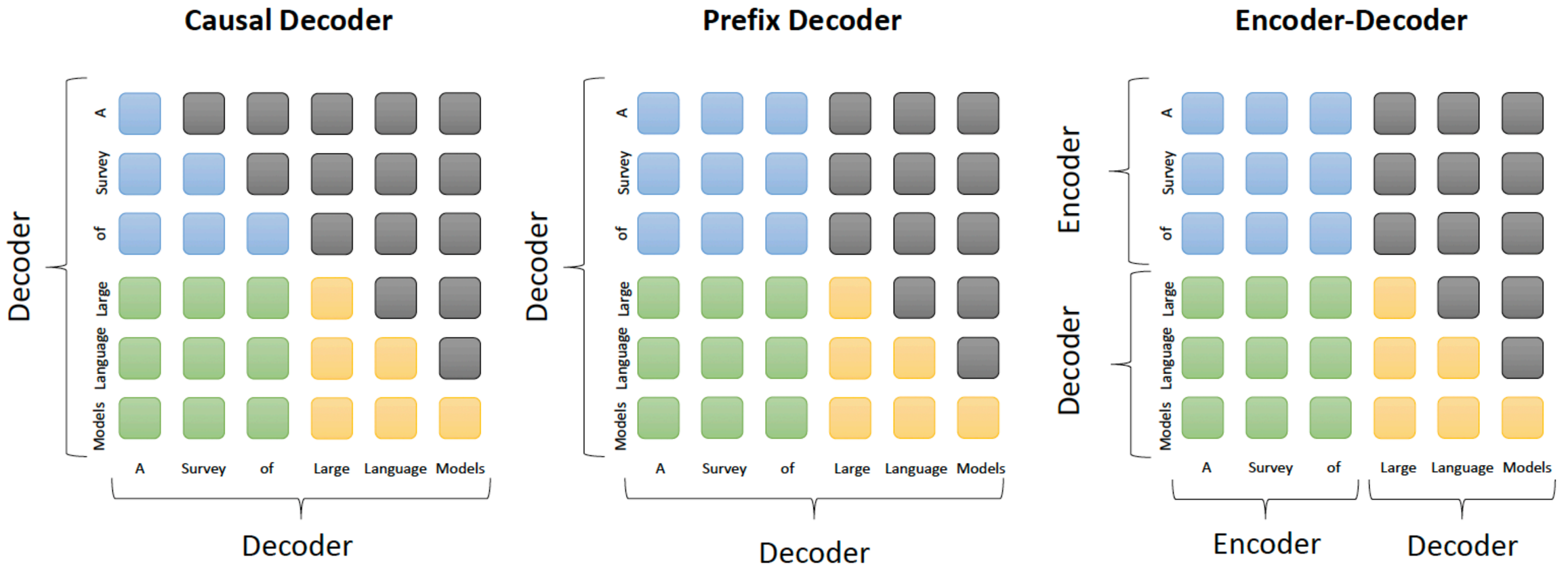

Fig. 7: A comparison of the attention patterns in three mainstream architectures. Here, the blue, green, yellow and grey rounded rectangles indicate the attention between prefix tokens, attention between prefix and target tokens, attention between target tokens, and masked attention respectively.