AI model capabilities & benchmark difficulties