Промпт-атаки на большие языковые модели (LLM)

Юшин С.В.

1 Задача

1.Вникнуть в тематику, разобраться с тем чем GenAI (в частности LLM) отличается от классических алгоритмов AI. Разобраться в уязвимостях специфичных для GenAI, как на этапе обучения, так и на этапе эксплуатации модели.

2.Подготовить список актуальных промпт-атак, реализуемых через окно диалога. Отдельное внимание уделить промпт-атакам в части кибербезопасности (КБ). «Использование LLM модели как инструмента подготовки кибератаки» - генерация вредоносного кода, социальная инженерия, фишинг и т.п. Поискать в открытых источниках и научных публикациях.

3.Попробовать провести промпт-атаку на LLM модели в свободном доступе. Оценить успешность атаки.

2 Оборудование

В работе используется: оптическая скамья, осветитель, два длиннофокусных объектива кювета с жидкостью, кварцевый излучатель с микрометрическим винтом, генератор У.З. волны, линза, вертикальная нить на рейтере, микроскоп.

3 Теоретические сведения

Источник света Л с помощью конденсора К проецируется на входную (коллиматорную) щель S. Входная щель ориентирована горизонтально и прикрыта красным светофильтром Ф. Коллиматорный объектив O_1 посылает параллельный пучок на кювету с водой С. Излучатель Q, погружённый в кювету, создаёт УЗ-волну. Вертикальное перемещение излучателя осуществляется винтом I, тонкая подача — лимбом II. При определённых положениях излучателя волна становится стоячей.

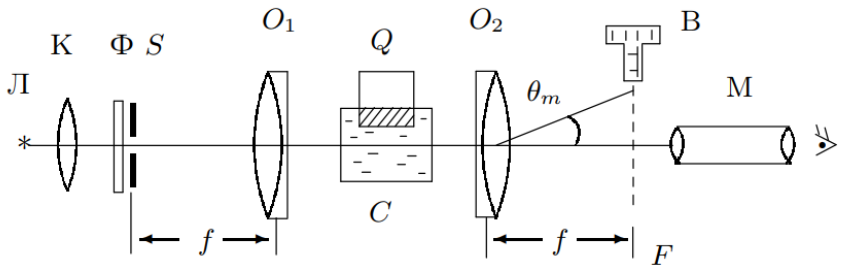

Рис. 1: Схема экспериментальной установки

Пусть фаза световых колебаний на передней поверхности жидкости равна нулю. Тогда на задней поверхности (т.е. в плоскости $z = 0$) она равна

$$\varphi = knL = \varphi_0(1 + m\cos\Omega x),$$

где L – толщина слоя жидкости в кювете, $k = 2\pi/\lambda$ – волновое число для света, λ – длина световой волны, $\varphi_0 = kn_0L$. Таким образом, в плоскости $z = 0$ фаза световых колебаний является периодической функцией координаты x, иными словами — УЗ-волна в жидкости создаёт фазовую дифракционную решётку.

Её функция пропускания:

$$t(x) = e^{im\cos\Omega x} \overset{m\ll 1}{\approx} 1 + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x}. \tag{1}$$

При освещении этой решётки плоской нормально падающей волной амплитуды a имеем за решёткой (при $z > 0$):

$$f(x,z) = ae^{ikz} + \frac{iam}{2}e^{i(\Omega x + \sqrt{k^2-\Omega^2}z)} + \frac{iam}{2}e^{i(-\Omega x + \sqrt{k^2-\Omega^2}z)}$$

При изучении дифракции методом тёмного поля будем удалять компоненту $f_0 = ae^{ikz}$ ставя проволочку в соответствующем месте фурье-плоскости. В этом состоит метод тёмного поля в изучении фазово-контрастных объектов.

При небольших амплитудах звуковой волны показатель преломления жидкости n меняется по закону

$$n = n_0(1 + m\cos\Omega x),$$

где Ω – волновое число УЗ волны, $m \ll 1$ – глубина модуляции УЗ волны.

В общем случае после прохождения через кювету световое поле представляет совокупность не трёх, а большого числа плоских волн, распространяющихся под углами, определяемыми условием

$$\Lambda\sin\theta_m = m\lambda,\ m \in \mathbb{Z}. \tag{2}$$

Каждая из этих волн соответствует одному из максимумов в дифракционной картине Фраунгофера. Определяя на опыте положение дифракционных максимумов различного порядка, можно по формуле ()**??**) найти длину Λ УЗ-волны и вычислить скорость v распространения ультразвуковых волн в жидкости, если известна частота ν колебаний кварцевого излучателя:

$$v = \Lambda\nu.$$

4 Результаты измерений и обработка данных

Исследование по дифракционной картине. Оценим *по порядку величины* скорость звука как удвоенное расстояние между наиболее чёткими дифракционными картинами:

$$n = 67 \text{ дел},$$

$$\lambda \approx 67 * 10 * 2 = 1340 \text{ мкм}.$$

Отсюда

$$v = \lambda * \nu \approx 1640 \text{ м/с}.$$

Эта величина не является точной, т. к. оценка проводилась по факту наибольшей видимости, поэтому подсчёт погрешностей не имеет смысла.

2*ν, МГц	a, мкм, в порядке n				
	-2	-1	0	+1	+2
1.4570	-344	-172	0	196	384
2.1515	—	-260	0	272	—
4.3971	—	-540	0	584	—

Таблица 1: Результаты измерений

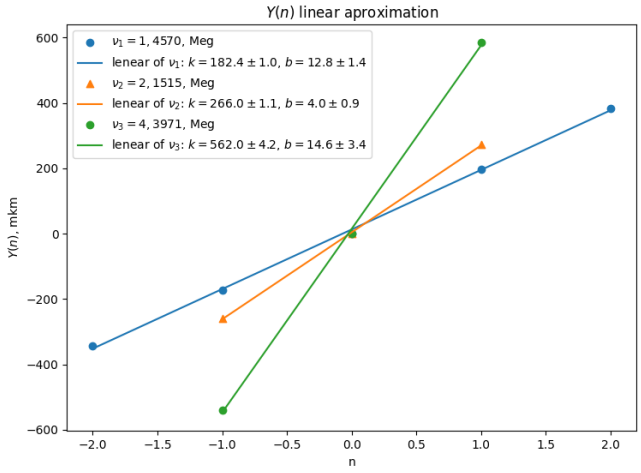

Рис. 2: Графики зависимости $Y = Y(n)$ (Y – мкм, n – б/р)

Определим положения дифракционных полос. Более 5 полос получить не удалось, т. к. генератор имеет низкую чувствительность ручки, а на высоких частотах ($\gtrsim 5$ МГц) выдаёт нестабильную частоту. Результаты в табл. **??**.

По результатам получим график на рис. **??**. В табл. **??** коэффициенты прямых и полученные по ним результаты из формулы

$$v = \nu m f \lambda / l_m = \nu f \lambda / k. \tag{3}$$

ν, МГц	1.4570	2.1515	4.3971
k	182 ± 1	266 ± 1	562 ± 4
v, м/с	1430 ± 20	1450 ± 20	1400 ± 30

Таблица 2: Результат расчёта скорости звука

Среднее значение:

$$v = 1430 \pm 50 \text{ м/с},$$

что близко к табличному значению $v = 1490$ м/с, но не сходится в пределах погрешности. Здесь случайная погрешность среднего взята по формуле среднеквадратичного отклонения (стандартной ошибки) и сложена с инструментальной по формуле $\sqrt{\sigma^2 + \delta^2}$.

Исследование методом тёмного поля. Найдём цену деления шкалы микроскопа через период сетки $h = 1$ мм. $n = 22$ дел/кл, т. е. 1 дел $= 45$ мкм

По формуле $\Lambda = 45$ мкм $* 2 * n/m$ найдём длину ультразвуковой волны. Резулаты измерений и расчётов в табл. **??**.

ν, МГц	1.7070	2.0866	4.2673	2.511	3.1819	3.8760
n, дел	65	44	43	48	47	30
m, линий	8	7	14	9	11	9
Λ, мкм	183	142	69	120	94	75

Таблица 3: Результат измерения длин волны

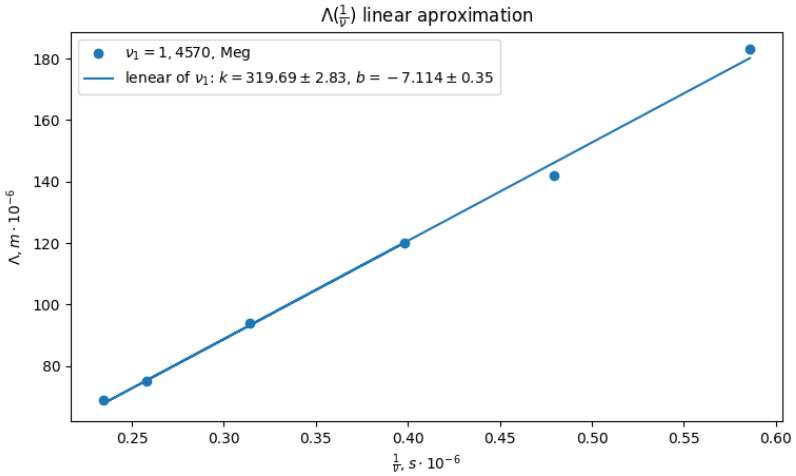

Рис. 3: Зависимость $\Lambda = \Lambda(1/\nu)$

Отсюда по графику на рис. **??** найдём скорость звука в жидкости:

$$\Lambda = k/\nu + b, k = 319 \pm 3 \text{ м/с}$$

$$v = 4.5 * k = 1436 \pm 15 \text{ м/с},$$

что согласуется с полученными ранее результатами.

Качественные наблюдения. Закрывая ненулевые максимумы получаем равномерную засветку, так как интенсивность нулевого максимума многократно превышает интенсивность ненулевых.

5 Вывод

Удалось с неплохой точностью измерить скорость звука в воде используя волны сжатие-разряжение как синусоидальную решётку: $v = (1430 \pm 50)\ m/s$, что лежит в пределах двух погрешностей от табличного значения: $v = 1490\ m/s$

Также, была изучена дифракция света на такой акустической решётке; Был применён и изучен метод тёмного поля в наблюдении фазовых объектов, и найдена с помощбю данного метода скорость звука в воде:$v = (1436 \pm 15)\ m/s$, что хорошо согласуется с эксперемынтальным результатом, полученным ранее.